ABSTRACT

Title of Document: POST-TRANSCRIPTIONAL REGULATION IN THE DROSOPHILA SEX DETERMINATION PATHWAY

David M. Sturgill, Ph.D., 2012

Directed By: Brian Oliver, NIDDK, NIH

Sexually reproducing organisms produce two very different phenotypes (males and females), by differential deployment of essentially the same gene content. This dimorphism provides an excellent model to study how transcriptomes are differentially regulated, which is one of the central problems of biology. The core sex determination pathway of Drosophila is a well described cascade of transcriptional and post-transcriptional regulation, but knowledge of the downstream components is largely incomplete.

High throughput technologies have provided great advances in understanding transcriptome regulation, but limits of the technology have lead to a focus on whole gene expression measurements, rather than post-transcriptional regulation. RNA-Seq experiments, in which transcripts are converted to cDNA and sequenced, allow the resolution and quantification of alternative transcript isoforms, potentially elucidating the post-transcriptional network. However, methods to analyze splicing are underdeveloped, and challenges in transcript assembly and quantification remain unresolved.

This work describes the development of the Splicing Analysis Kit (Spanki) as a fast, open source, suite of tools that uses simulations based on real RNA-Seq data to characterize errors in a given dataset, and user tunable filters that minimize those errors. Spanki quantifies splicing differences in transcripts from the same loci within a sample, as well as between samples by using only those reads that directly assay splicing events (junction spanning reads). Despite the reliance on a fraction of the total data, sequencing depth typically generated in an RNA-Seq experiment is sufficient to identify differentially regulated splicing, and error profiles are superior. I demonstrate that this computational approach outperforms several commonly used approaches in an analysis of sex-differential splicing in Drosophila heads.

Next I examine the effects of disrupting post-transcriptional regulation in Drosophila heads. I apply the Spanki software to analyze RNA-Seq data for mutant lines of two post-transcriptional regulators: *Darkener of apricot* (*Doa)* and *found in neurons* (*fne)*. *Doa*, a serine-threonine kinase, regulates splicing by phosphorylating SR proteins, vital components of the splicing machinery. *Found in neurons* (f*ne*) binds to transcripts and is involved in RNA metabolism. I demonstrate sex-differences in response to disruption of post-transcriptional regulation, and hypothesize that they are informative of sex-differentiation pathways.

Finally, I examine the conservation of splicing regulation within the Drosophila lineage. I show that junction based splicing analysis is effective in making interspecific comparisons without the need for complete transcript models. I use these results to demonstrate the conservation of sex-differential splicing across 40 million years of evolution in 15 species in the Drosophila genus.

POST-TRANSCRIPTIONAL REGULATION IN THE DROSOPHILA SEX DETERMINATION PATHWAY

By

David M. Sturgill

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Professor Steve Mount,Chair
Dr. Brian Oliver
Professor Eric Haag
Professor Najib El-Sayed
Professor Carlos Machado

Preface

Portions of this dissertation have either been published in peer-reviewed journals or are in preparation for submission. I owe a great debt to each of my co-authors for these manuscripts. Chapter 2 describes the toolkit I built to analyze splicing from RNA-Seq data. I conceived and implemented the computational work for this analysis. This chapter is based on a manuscript (preparing for resubmission at time of writing), which is the combined work of myself, and our collaborators. Chapter 3 describes an analysis of mutant samples generated by our collaborations. Chapter 4 describes comparative analysis of Drosophila species.

Drosophila species results are partially described in this modENCODE publication:

The developmental transcriptome of Drosophila melanogaster. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE. Nature. 2011 Mar 24;471(7339):473-9. Epub 2010 Dec 22.

Additional Drosophila species data come from samples prepared previously and described here:

Constraint and turnover in sex-biased gene expression in the genus Drosophila.
Zhang Y, David Sturgill, Parisi M, and Oliver B.
Nature. 2007 Nov 8;450(7167):233-7.

Acknowledgements

This work would not have been possible without contributions from many people I have met during my time at the NIH and the University of Maryland. My experience has shown me that good science is never done in a vacuum, and collaboration is the norm. Computational scientists often have to rely on others to produce quality primary data to analyze, and I am grateful and fortunate to have had many talented collaborators and colleagues.

Chapter 1 provides an overview of the biological questions my research addresses, and a general introduction to the analytical approach. My background knowledge of sex determination in Drosophila is thanks to my advisor Brian Oliver, for always reminding me to focus on the biological question at hand, and for sharing his expertise in the subject. Portions of this section cite our previous work on comparative expression analysis with microarray data. For those manuscripts, I worked together with Yu Zhang. They were the product of working through many long hours and a lost Christmas vacation, and I owe Yu a great debt. Also mentioned in the section is a comparison of dosage compensation across mammals, worms and flies; and I contributed to the Drosophila component. I am thankful to Xinxian Deng and Christine Disteche for pulling that manuscript together.

Chapter 2 describes the software I built called Spanki (Splicing Analysis Kit). I quickly learned that computational tools lag behind every new assay technology, and found that I had to build tools from the ground up to accomplish the analysis I wanted. I am most thankful for our collaborators Leonard Rabinow and Marie-Laure

Samson who generated the samples that were used as a test case for analysis, and for John Malone who completed the sequencing and contributed a great deal to the early processing and analysis of these data. I also thank Yunpo Zhao and Xia Sun for designing and performing PCR experiments for validation. Marie-Laure also contributed significantly to the development of the software, by always keeping a keen eye on details and finding places where the algorithm could be improved.

For the software development, I am grateful to Ryan Dale for his valuable advice in packaging it up and distributing it. He saved me a great deal of traversing up a learning curve. I also owe a lot to Cole Trapnell, who was generous with his time to discuss things with me during a visit to College Park, and whose software was pioneering work for RNA-Seq analysis. His software is a great example of the benefit of open source, and I studied Tophat as a well-engineered tool to take as an example of how to code. I also owe gratitude to the work of Mike Duff, Jane Landolin, and Angela Brooks, whose excellent computational work contributed greatly to the modENCODE project.

Participation in the modENCODE project has been a great experience and a valuable exercise in collaborative science. I learned a great deal from working with the fly transcriptome Analysis Working Group (AWG) headed by Sue Celniker, and I thank them for their advice and comments about Spanki.

Chapter 3 describes analysis results for mutants of post-transcriptional regulators, and again I owe a debt to Lenny, Marie-Laure and John, without whom these data would not exist. I owe Lenny and Marie-Laure another round of thanks, for their patience, encouragement, and advice on many conference calls. I learned a great deal about manuscript writing and about the biology of these mutants from them. The work on this project further reinforced in me the need for computational scientists to understand the biology of the samples they are analyzing and the technical details of the data they are processing.

Chapter 4 provides an analysis of differential expression in multiple Drosophila species. For these data, again I thank Yu Zhang and Mike Parisi, who helped prepare some of the original samples. I particularly thank Mike Parisi, who had the patience to walk me through generating RNA samples, which was a great learning experience for me. Additional samples and RNA-Seq experiments were done by John Malone, Nicolas Mattiuzo, and Carlo Artieri, and I owe a great debt to them. Generating and analyzing these data was a monumental task that took a team effort and a great deal of time. I also want to acknowledge Zhen-Xia Chen, who has been provided valuable computational support.

Aside from these specific chapters, I want to acknowledge many people who have contributed to my development as a scientist over the years. First, there have been many colleagues that have come though the Oliver Lab, which have all left me with positive experiences: Mike Parisi, Sandra Farkas, Vaijayanti Gupta, Yu Zhang, Revital Bronstein, Rasika Kalamegham, Jamileh Jemison, Leonie Hempel, John Smith, Nello Cerrato, Lee Hangnoh,, Renhua Li, Emily Clough, John Malone, Carlo Artieri, Nico Mattiuzo, Hina Sultana, Zhen-Xia Chen, Allen Gibbs, and Tim Westwood.

There are many investigators that the NIH I have met and been fortunate to work with who I would like to acknowledge. I thank Chuck Vinson and Peter

Fitzgerald for their collaboration on core-promoter analysis. This work was a great scientific and personal experience, and I particularly thank Chuck for his encouragement and advice over several years. I also thank the members of the multiple endocrine neoplasia (MEN) analysis group, particularly Francis Collins and Stephen Marx for the great experience of sitting in and contributing to their science. It was one of the highlights of my stay at the NIH and I still remember the praise and feedback for my contribution.

None of this would have been possible without the help of Brian Oliver, who brought me to the NIH in 2004 as a bioinformatician. His relationships are what made these collaborations possible. I owe him a great deal of gratitude for his rigor and for making me a better scientist.

There were many people at College Park who I also owe thanks. First is Steve Mount, who shared his knowledge and ideas about splicing, and who has been essential to guiding me through the Ph.D. program. I want to acknowledge others who served on my committee: Steven Salzberg, who gave valuable advice and served on my committee, and allowed me to work with and learn from his students; Cristian Castillo-Davis, who forced me to be rigorous in thinking about evolution; Eric Haag, who provided great insight and intellectual support, as well as great advice; Najib El-Sayed, who gave great feedback and encouragement; and Carlos Machado, who provided helpful discussions and contributed his expertise about Drosophila species.

I also owe thanks to Cynthia Gibas, my Masters degree mentor at Virginia Tech. Her book was what initially inspired my to enter the field, and her mentorship served as a great example of good academic work. Her guidance, along with Jennifer Weller, provided a launching point for me.

On a personal level, I owe a huge debt of gratitude to my friends and family, above all else my wife Julieta. She has been my rock and confidant, and supported me through many stressful times while I completed this work. I also want to thank our dog Smokey, who was my constant companion through many hours of writing and reading manuscripts at home, and during stress-relieving walks in the woods.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Section 1.1 Regulation of the transcriptome

Complex multicellular organisms such as humans are composed of a variety of vastly different cell types. With few exceptions, each cell type within an organism includes the same genetic instructions. A major gap in our knowledge of biology is a thorough understanding of how different cells produce different phenotypes from the same genotype.

Since each cell contains the same genetic information, the difference must arise in how this information is decoded and used. There are many means for regulating how the genetic information is used, either by modulating when and how much genes get transcribed (transcriptional regulation) and by altering how transcripts are processed to generate protein (post-transcriptional regulation). The term used for the complete set of transcripts produced by the cell is the "transcriptome" (Figure 1.1). Each cell may have identical genomes, but very different transcriptomes. These differences can consist of different relative amounts of the same transcripts, and/or qualitatively different transcripts that are processed and altered differently. All of these differences can arise by differential deployment of regulatory mechanisms by the cell.

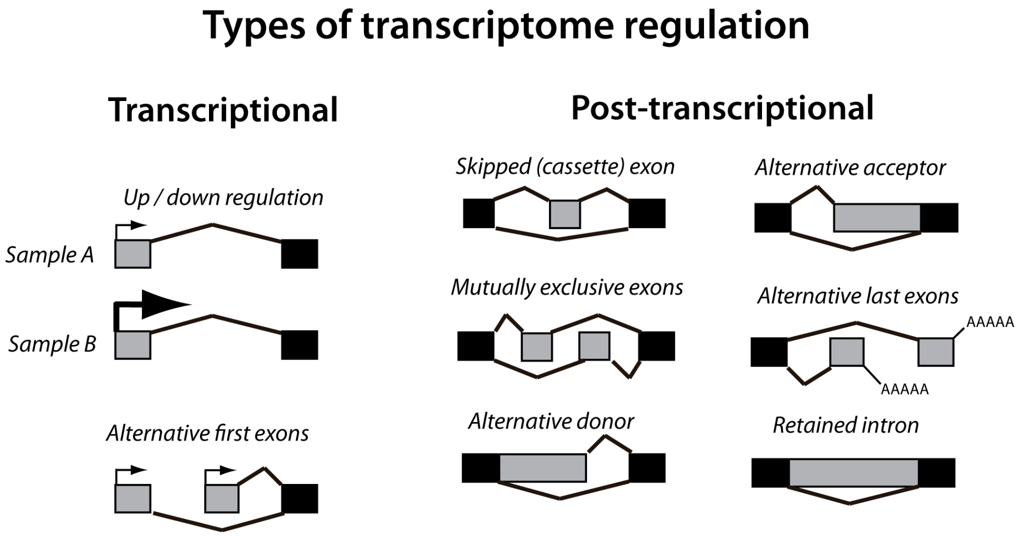


Figure 1.1: Types of transcriptome regulation

1.1.1 Transcriptional regulation

Genes must be transcribed to RNA for their information to be used. One way this can be regulated is to modulate quantitatively how much transcript is produced (Figure 1.1). This is accomplished by regulating transcriptional activity at promoters (Lee and Young 2000). Quantitative regulation involves up- or down-regulation at promoters to alter transcriptional activity, and involves both trans-acting factors and cis-regulatory elements. Transcription can proceed along tightly regulated developmental programs, following a cis-regulatory code (FitzGerald, Sturgill et al. 2006; Sorge, Ha et al. 2012).

Transcription from alternative promoters of the same gene can also be regulated to produce different transcripts, and this is widespread in humans (Davuluri, Suzuki et al. 2008; Singer, Wu et al. 2008). Although this is regulated at the level of transcription, alternative first exon events (AFEs) are commonly included as a form of alternative splicing (Black 2003), since it leads to isoform variants that have a splicing difference.

1.1.2 Post-transcriptional regulation

The other major way cells can shape their transcriptomes is after transcription, by alternative splicing (Black 2003, Figure 1.1). In this process, eukaryotic organisms can produce multiple distinct proteins from one type of primary transcript. Alternative splicing generates different RNA molecules from identical primary transcripts, affecting protein diversity by creating diverse mRNA isoforms and modulating regulatory information in non-coding and untranslated regions in mRNAs (Black 2003). This process greatly increases the number of proteins that may be produced from a gene by combinatorial complexity.

Alternative splicing is widespread in eukaryotes, but high-throughput methods to provide a complete accounting of it have been lacking. Estimates suggest that 60-99% of genes are alternatively spliced in humans (Pan, Shai et al. 2008; Wang, Sandberg et al. 2008), but is less prominent in Drosophila (40% of annotated genes) (Graveley, Brooks et al. 2011).

Splicing can be regulated by biological context; such as between tissues, sex, and developmental stage; and is an important mechanism for the sexes to produce different RNA-output in different cellular and developmental contexts (Black 2003). Splicing is precisely regulated, and many human diseases, such as cystic fibrosis, are caused by errors in splicing or the presence of specific splice variants (Garcia-Blanco, Baraniak et al. 2004).

In Drosophila, alternative splicing is critical to the core sex determination pathway (Figure 1.2), but not all downstream targets of splicing regulation are known. One survey of sex-specific alternative transcription estimated that 11-24% of Drosophila genes are alternatively spliced in a sex-biased manner (McIntyre, Bono et al. 2006), although this study did not consider the entire genome. To date, there has not been a transcriptome-wide comparison of sex-differential splicing in Drosophila published.

Expressed sequence tag (EST) sequencing, and microarrays that target splice junctions and exons, have helped elucidate splicing, but have not provided a complete picture. High-throughput sequencing of transcripts (RNA-Seq) with short reads provides us with the resolution to capture sequence from all isoforms (Marioni, Mason et al. 2008; Mortazavi, Williams et al. 2008). However, computational methods to resolve relative abundances of isoforms in a sample, and making meaningful comparisons across samples, are currently not fully developed.

Section 1.2 Drosophila as a model of transcriptome regulation

Drosophila has been a model organism for over 100 years, and has well characterized genetic pathways, making it an excellent system to study transcriptome regulation (Bellen, Tong et al. 2010; Yamamoto 2010). The core sex determination pathway, which describes the regulatory mechanisms by which the initial developmental decision about gender is initiated and maintained, is an example of a particularly well characterized pathway (Venables, Tazi et al. 2011, Figure 1.2).

1.2.1 The sex determination pathway

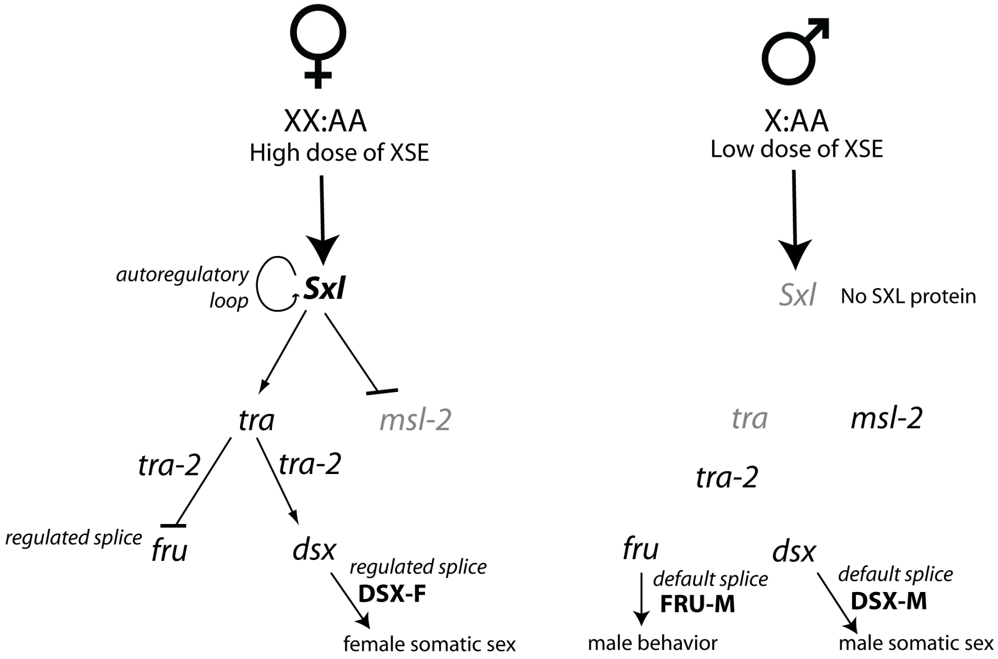


Figure 1.2: Somatic sex determination in Drosophila. Adapted from B. Baker: (http://cmgm.stanford.edu/devbio/baker/Hierarchy.htm) and (Verhulst, van de Zande et al. 2010).

The sex determination pathway in Drosophila is a regulatory cascade that receives a signal (expression dose of X-linked Signal Elements, or XSEs), and transmits this decision about sex to direct differentiation in the soma. Most relevant to this study, the Drosophila sex determination hierarchy is also classical model of regulated alternative splicing. Transcripts of three members of this hierarchy, *Sex-lethal* (*Sxl*), *transformer* (*tra*), and *male specific lethal 2* (*msl-2*) are broadly expressed. The two terminal members of the hierarchy, *doublesex* (*dsx*) and *fruitless* (*fru*), are transcription factors that regulate expression of downstream targets. All of these sex determination genes are regulated by sex-differential alternative splicing

(Venables, Tazi et al. 2011, Figure 1.3). A close examination of this pathway reveals the diverse functional elements and mechanisms at play, including transcriptional enhancers, polypyrimidine tracts, in-frame stop codons, and exonic splicing enhancers.

The cascade initiates early in embryogenesis, with a promoter of the *Sxl* gene responding to a signal of X-chromosome dose via expression of the XSEs (*unpaired* , *runt*, *sisA* and *scute*) (Sanchez, Granadino et al. 1994; Salz and Erickson). In females, the X chromosome is present in two copies, and the XSEs are expressed at a level that activates the *Sxl* early promoter. In males, XSE expression does not reach the threshold level, and the *Sxl* early promoter remains inactive (Salz and Erickson). Expression from this early promoter of *Sxl* represents the initial decision about sex by the organism.

This decision about sex is maintained by an autoregulatory feedback loop. In later developmental stages, constitutive transcription of *Sxl* occurs from a different promoter (Cline 1984). The absence of pre-existing SXL protein leads to default splicing of the *Sxl* pre-mRNA, which contains a premature termination codon (PTC) . In females, pre-existing SXL binds to the pre-mRNA, leading to a splicing variant that leads to more functional SXL (Cline 1984; Bell, Horabin et al. 1991).

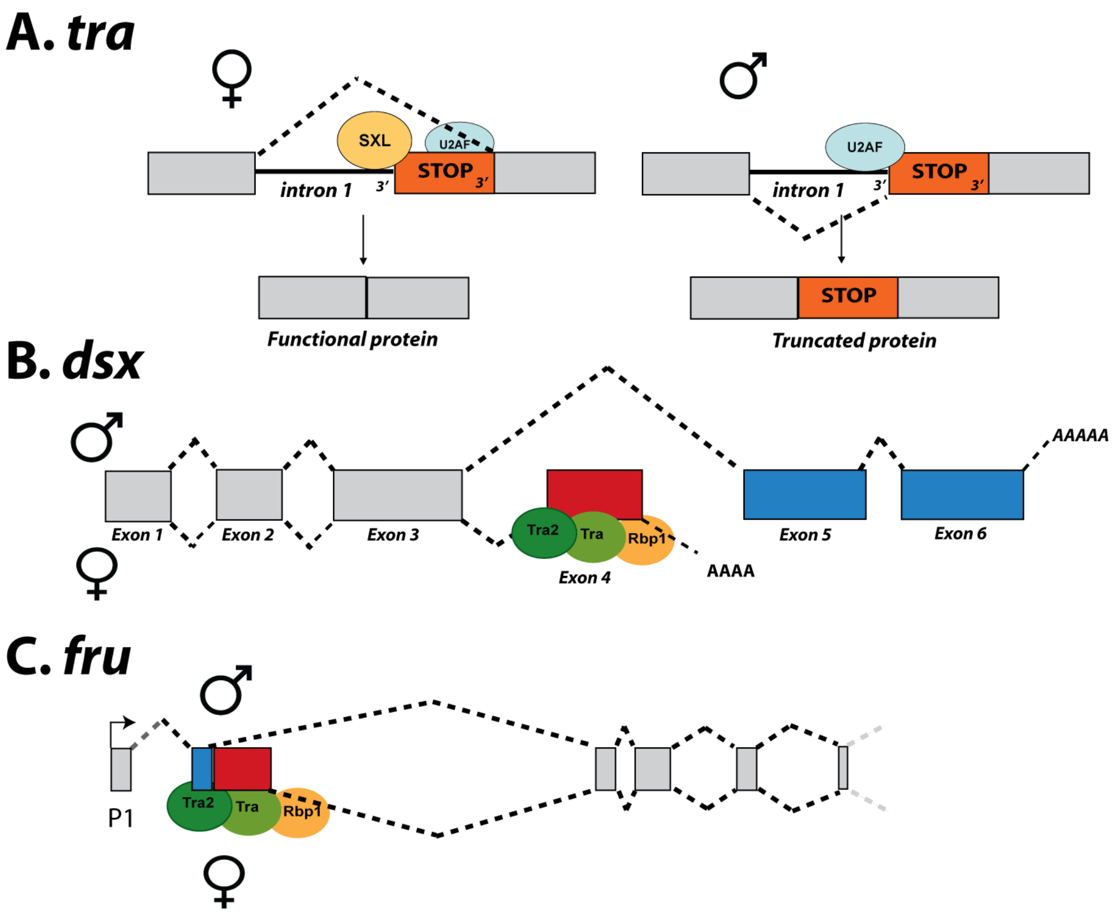


Figure 1.3: Sex-differential splicing in the sex determination hierarchy. (A) Splice variant in males leads to non-functional protein. (B) Two functional variants of DSX., determined by presence of functional TRA. (C) Differential splicing at an alternative donor, in transcripts of *fru* from the P1 promoter.

The presence of SXL in females also turns off dosage compensation, by translational control of *msl-2* pre-RNA (Bashaw and Baker 1997; Beckmann, Grskovic et al. 2005). SXL binds to a polypyrimidine tract in an intron of the 3' UTR of *msl-2* pre-mRNA , causing this intron to be retained. Bound SXL then affects recruitment of 43s ribosomal preinitiation complexes. This prevents MSL-2 translation from occurring, which ensures X-chromosome dosage compensation does not occur in females.

Binding of SXL to pre-mRNA of *tra* leads to the recognition of an alternative 3' acceptor site and eventual translation (Sosnowski, Belote et al. 1989). Without

SXL in males, a default splicing of *tra* occurs that leads to a transcript with a PTC (Figure 1.3A). TRA protein is an essential splicing regulator that guides the sex-specific splicing of downstream transcription factors (Sosnowski, Belote et al. 1989; Salz and Erickson, Figure 1.2).

The culmination of this initial sex-determination hierarchy is the differential splicing of transcripts of *dsx* and *fru*, both encoding DNA-binding transcription factors (Lynch and Maniatis 1996; Demir and Dickson 2005). In females, with a functional TRA (and along with TRA-2), the transcript from *dsx* is spliced to DSX-F (including exon 4, Figure 1.3B). The 3' splice site for exon 4 contains a poor polypyrimidine tract that contains several purines, making a splicing enhancer (TRA) necessary (Burtis and Baker 1989). Without TRA, the default splicing is to skip exon 4 and produce DSX-M. A different functional protein is then produced in each sex from these different processed transcripts. Both of these proteins have identical DNA-binding domains but differ in their carboxy termini (Burtis and Baker 1989; Shukla and Nagaraju 2010, Figure 1.3).

fruitless (*fru*) transcripts are also targeted by TRA and TRA-2, producing sex-specific isoforms, but only the male variant from the an upstream promoter (P1) is translated (Siwicki and Kravitz 2009, Figure 1.3C). Like DSX, the FRU protein is a transcription factor, with properties of the zinc-finger family of DNA binding proteins. *fruitless* is active in the central nervous system, and it is an example of a current focus of research to identify the genetic basis for reproductive behaviors (Moehring, Li et al. 2004; Vosshall 2007). The male-specific splice variant of *fru* was shown to direct male mating behavior, when exogenous expression of the male-specific splice variant of *fru* induced male behavior in female flies (Demir and Dickson 2005). The downstream targets of *fru* that produce this behavioral phenotype remain to be discovered.

1.2.2 Drosophila as a system for evolutionary divergence

Drosophila is an ideal model for evolutionary research. There are many Drosophila species distributed globally with diverse morphologies and living in tropical, urban, and desert environments. They are easily culturable with short generation times, and are excellent for conducting genetic experiments. The lineage spans an estimated 40 million years (Clark, Eisen et al. 2007); for perspective, the estimated divergence time of humans from chimpanzees is 6 million years and from New world monkeys is 33 million years (Glazko and Nei 2003). Reference genomes have been sequenced for species from a range of close and distant time scales (Figure 1.4). All these characteristics make the Drosophila genus an excellent model system to study phylogenetic divergence.

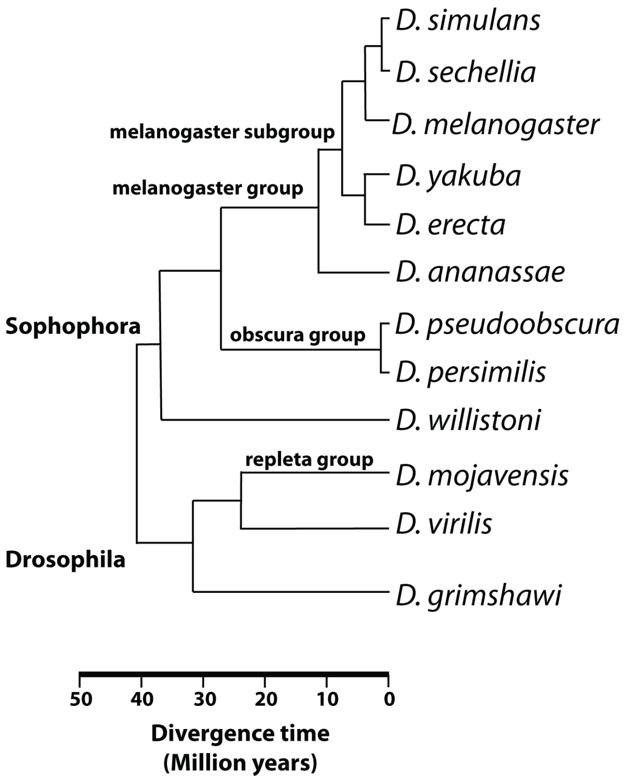


Figure 1.4: The phylogeny of sequenced species of Drosophila (Clark, Eisen et al. 2007; McQuilton, St Pierre et al. 2012).

<u>*Section 1.3 Divergence of the transcriptome*</u>

Species phenotypes diverge over evolutionary time, driven by mutation that alters genomically encoded information. The most familiar way this occurs is by the alteration of coding sequence to change the encoded protein. However, changes in gene regulation are also hypothesized to have an important role in species divergence (Romero, Ruvinsky et al. 2012).

Comparative expression analysis is a rapidly growing field that has begun to explore gene expression variation over time. The rise of high-throughput methods has recently allowed the comparisons of whole transcriptomes of divergent species. These studies have shown that selection pressure acts on gene regulation, and have shown patterns of lineage-specific adaptive change in expression (Romero, Ruvinsky et al. 2012). Results in Drosophila have shown that sex-differential gene expression diverges over evolutionary time, so that one can generate a phylogenetic tree using measures of expression divergence that mirrors a phylogeny based on sequence divergence (Zhang, Sturgill et al. 2007). At the same time, expression divergence patterns for subsets of genes may appear stochastic and not directly correlated with sequence divergence (Zhang, Sturgill et al. 2007). These results suggest that variation in gene deployment between species is significant, and that selection acts not just on coding sequence but also on transcriptional regulation.

One major challenge of this growing field is to distinguish changes in expression that are adaptive from random variation and drift. In coding sequence evolution, one can compare the rate of change at synonymous and non-synonymous sites to infer the type of selection taking place (Hurst 2002), and there is no analogous test for gene expression. A natural place to look for adaptive divergence is in sexually dimorphic features, since sexual selection for advantageous adaptations is the engine that drives evolution.

1.3.1 Selection pressure on dimorphic features

Sexual selection has been known to be major driving force of evolutionary divergence for nearly 150 years (Darwin 1871). Beneficial adaptations that enhance mating effectiveness have a selective advantage, and propagate within a population (Clutton-Brock 2007). Positive selection on these traits can act on coding sequence, cis-regulatory elements for transcription, or post-transcriptional regulatory elements (Xing and Lee 2006).

Traits that influence reproductive success are prevalent in the Drosophila lineage. One example is sperm tail length, which is highly variable in the Drosophila lineage, and hypothesized to correlate with reproductive success (Joly, Korol et al. 2004). Selection pressure on sperm tail length is substantial enough that one species of fly has evolved the longest sperm cell in the animal kingdom, 300 times longer than human sperm (Pitnick, Spicer et al. 1995).

One important force constraining transcriptome divergence is sexual antagonism, where gene products are beneficial to one sex and detrimental to the other (Innocenti and Morrow 2010). In these cases of intra-locus conflict, the organism must balance the relative costs and benefits to the two sexes, or find some means to regulate expression sex-specifically. This regulation can be transcriptional, or post-transcriptional via splicing. Splicing divergence is a particularly appealing mechanism to develop sexually dimorphic features, which may benefit sex over the other. Alternative splicing may enable males and females to generate different proteins from the same gene without generating sexually antagonistic effects.

1.3.2 Splicing divergence

Conservation of splicing patterns is an important component of stabilizing selection pressure acting on genomes, and divergence of splicing patterns is hypothesized to be a major mechanism of generating phenotypic complexity (Boue, Letunic et al. 2003; Xing and Lee 2006). By the generation of new exons and splicing patterns, and allowing exons within a gene to evolve under different selective pressures, it allows a "trial-and-error" approach to generating new gene content (Boue, Letunic et al. 2003). Estimates from human / mouse ortholog pairs suggest that about half have species-specific isoforms (Modrek and Lee 2003), and a comparison among Dipteran flies (Malko, Makeev et al. 2006) has also revealed extensive isoform species specificity. In Drosophila, one microarray study examined 417 genes, and suggested that most sex-differential splicing in Drosophila is conserved across species (Telonis-Scott, Kopp et al. 2009), but a more comprehensive investigation may uncover many species-specific isoforms.

There are multiple mechanisms by which splicing may diverge, including changes to cis-regulatory elements and trans-factors. A major path for the evolution of new exons is by duplication. Kondrashov *et al.* suggest that tandem exon duplication followed by alternative splicing has had an important role in expanding the functional and regulatory diversity of the genes involved (Kondrashov and Koonin 2001; Xing and Lee 2006). New splicing patterns can arise when entire genes duplicate. When genes duplicate, each copy may diverge to perform different functions through subfunctionalization. This divergence can manifest as differences in splicing between the duplicates. Kopelman et al. show that singleton genes (with no duplicates) have more splice forms than those in multigene families, suggesting that singletons are more likely to use alternative splicing than genes that have undergone duplication (Kopelman, Lancet et al. 2005). Their results are consistent with the idea that an alternatively spliced exon may serve as an 'internal paralog' of a gene. This inverse relationship between duplication and splicing can be explained by a" balanced fulfillment of a requirement for diversification through either of the two mechanisms" (Kopelman, Lancet et al. 2005).

Splicing inclusion / exclusion patterns of orthologous exons may also diverge by transition (Keren, Lev-Maor et al. 2010) where a constitutive exon can become alternative by mutation in intronic or exon splicing enhancers. If this is the case, conservation within introns should also reveal important regulatory sequence. For example, a comparison between human and mouse showed that the flanking regions of alternatively spliced exons are significantly more conserved than that of constitutive exons (Sorek and Ast 2003). Higher conservation proximal to introns regulated by tissue has also been observed (Sugnet, Srinivasan et al. 2006). In this study, Sugnet et al. also looked in more detail at the conservation pattern of one family of genes, and showed that intron sequences have diverged between these paralogs, even though they share a regulatory pattern. In this unusual case, orthologous members of the family had highly conserved intron sequence, which suggests that the intron sequences are important to regulation. This example demonstrates that a simple positive correlation between splicing regulation and intron sequence conservation may not always be evident.

1.3.3 Evolution of sex determination

Sexual dimorphism is an ancient feature of the eukaryotic lineage. Sex determination mechanisms however exhibit diversity across phyla despite using common conserved components (DM domain genes) (Haag and Doty 2005). Closely linked to sex determination is dosage compensation, which also may use different mechanisms to achieve the same goal in different taxa (Deng, Hiatt et al. 2011).

Even within insects, there is a great variety of sex-determination systems (Sanchez 2008). In Drosophila, it is hypothesized that the sex-determination hierarchy evolved from a "bottom-up" pattern, where the terminal ends of the sex-determination hierarchy (*dsx* and *fru*) are more conserved than their upstream regulators. DM domain proteins (such as DSX and FRU) proteins, are broadly conserved across eukaryotes from worms to humans. *Sxl* and *tra* on the other hand, are more evolutionarily labile. Even within flies, it was observed that *Sxl*, the master regulator, does not have a sex-determining role in Houseflies (Meise, Hilfiker-Kleiner et al. 1998) and Mediterranean fruit fly (Saccone, Peluso et al. 1998).

In Drosophila, the leading hypothesis is that the *tra – dsx* axis was the primary regulator of sex determination in a distant common ancestor (Shearman 2002). *Sxl* was recruited to become the master regular of sex determination later, after their divergence from other Dipterans, and is much more evolutionarily labile (Traut, Niimi et al. 2006; Sanchez 2008). A closer examination of each of the pathway's components demonstrates the complexity of their divergence patterns.

1.3.4 Conservation of the sex-determination pathway components

Sex-lethal (Sxl)

Sxl orthologs have been defined in each of the 12 sequenced Drosophila species (Clark, Eisen et al. 2007). However, within the genus, Sxl gene models and regulation patterns are not clearly conserved. In *D. virilis*, the *Sxl* ortholog has been shown have a different exon-intron structure than *D. melanogaster.* Males of *D. virilis* also produce abundant SXL protein (Cline, Dorsett et al. 2010).

Although *Sxl* orthologs are present in non-Drosophila flies (Meise, Hilfiker-Kleiner et al. 1998; Saccone, Peluso et al. 1998), their role in sex determination is not conserved. Orthologs have also been described in the Lepidopteran *Bombyx mori*, but it is not sex-differentially spliced (Sanchez 2008). In non-Dipteran insects, *Sxl* serves a non sex-specific regulatory role, with conserved RNA-binding domains but without sex-differential splicing. This suggests that *Sxl* was co-opted from a general regulatory function early in the Dipteran lineage to serve as the master sex-determination switch (Sanchez 2008).

DM domain genes: ***doublesex*** **(*dsx*)** and ***fruitless*** **(*fru*)**

The DM domain is zinc-finger DNA binding module that is involved in sex determining mechanisms in diverse lineages from worms to humans (Murphy, Zarkower et al. 2007; Matson and Zarkower 2012). DSX and FRU are both members of this protein family.

doublesex is conserved in all sequenced Drosophila species, with a lower rate of change than the other sex determination components (Mullon, Pomiankowski et al. 2012). The genomic binding sites for DSX target genes have also been shown to be conserved across Drosophila species (Luo, Shi et al. 2011). In other species of insects, orthologs have been described in mosquito, and sex-differential regulation is conserved in *Aedes* and *Anopheles* (Salvemini, Mauro et al. 2011).

FRU also contains a DM domain, and is involved in generating sex differences in behavior. Orthologs have been defined in non-*melanogaster* Drosophilids (*D. sechelia*, *D. pseudoobscura*, *D. mojavensis*, *D. erecta*, *D. simulans*), and mosquito (*Culex quinquefasciatus*) (OrthoDB, (Waterhouse, Zdobnov et al. 2011)). However, high-confidence alignments have not been possible in current genomic assemblies of the other Drosophila species to define orthologs (Mullon, Pomiankowski et al. 2012). In experiments before the sequencing of these genomes, orthologs with conserved molecular structure were identified in *D. simulans*, *D. yakuba*, *D. pseudoobscura*, *D. virilis*, and *D. suzuki* (Billeter, Goodwin et al. 2002); and also in *A. gambiae* and *Tribolium* (Gailey, Billeter et al. 2006). In all cases, sex-specific splicing was conserved (Sanchez 2008).

Section 1.4 High-throughput methods for studying the transcriptome

Experimental methods to analyze gene expression have been in use for decades and are now routine. These methods can be used to interrogate transcripts (PCR, Northern blots) or proteins (Western blots). Analysis of entire transcriptomes requires more high-throughput methods. This throughput has been available for the past 20 years, and has lead to great advances in biological knowledge.

1.4.1 Microarrays

Whole transcriptome studies became possible about 20 years ago with the advent of microarrays (Schena, Shalon et al. 1995). In this method, oligonucleotide probes are covalently bound to a glass slide, and fluorescently labeled RNA is washed over the slide. The RNA hybridizes to the oligonucleotide probes, and total fluorescence is quantified with a laser scanner. The transcript abundance in a source pool of RNA is measured by fluorescence intensity. To detect splicing, specialized microarrays can be designed that target exons or splice junctions (Cuperlovic-Culf, Belacel et al. 2006).

Microarray technology has lead to great advances in understanding of how transcriptomes are regulated, by comparing transcription across tissues, developmental stages, disease states, and species. Microarray experiments are relatively inexpensive and easy to perform, allowing the generation of data for many samples. Microarray data have become a great resource for biologists to interrogate, with more than half a million experiments archived in the Gene Expression Omnibus (GEO) (Malone and Oliver 2011).

Although the wealth of knowledge from microarrays has been great, there are some areas where this technology is lacking. Prior knowledge of the transcriptome being targeted is required, since probes for microarrays must be designed. This limits their use to species for which good genomic reference sequence and annotation already exist. Oligonucleotide probes must also be thoughtfully designed to hybridize to their target effectively and specifically, which limits the transcript regions that much be targeted.

For these reasons, microarrays are not effective for comprehensive characterization of pools of RNA. They are limited to only detect what they are designed to detect, and therefore can not characterize the transcriptome of underannotated species, identify unknown transcribed features, or detect unknown aberrant events, such as splicing errors. For tasks such as this, new next-generation sequencing technology is critical (Figure 1.5).

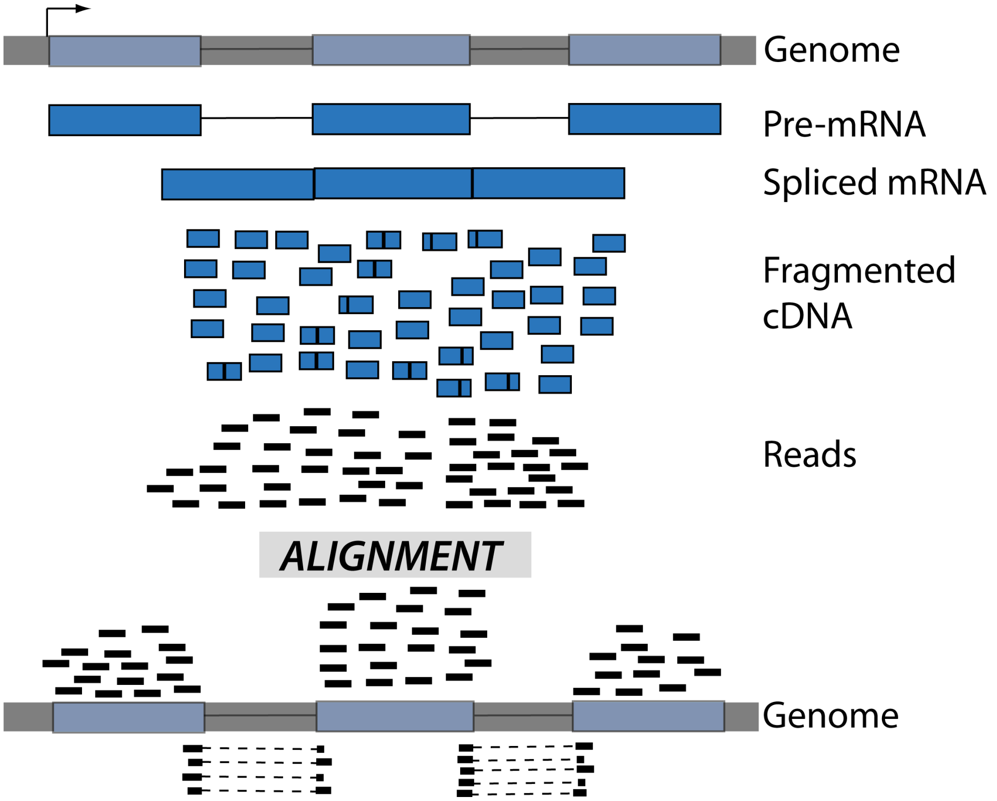


Figure 1.5: Illustration of RNA-Seq

1.4.2 RNA-Seq

RNA-Seq allows the analysis of whole transcriptomes at a resolution that was not possible with previous technology (Blencowe, Ahmad et al. 2009). In a typical RNA-Seq experiment, Poly-A+ transcripts are enriched from a pool of RNA, from which cDNA is generated, amplified, and sequenced (Oshlack, Robinson et al. 2010). The reads that are produced contain sequence that originated from transcribed exons, some of which contain sequence from two exons spliced together (Figure 1.5). The first step of a typical analysis entails aligning reads to a reference genome, to help infer the transcript molecule from which the read originated. From these results, downstream analysis involves estimation of relative abundances of transcribed and processed features (Oshlack, Robinson et al. 2010; Martin and Wang 2011).

Despite this added resolution, there are important sources of ambiguity, bias, and noise in RNA-Seq data that have made accurate abundance estimation of isoforms difficult in practice (Malone and Oliver 2011). These problems arise at multiple steps in an RNA-Seq experiment. For example, at the biological level, introns retained in incompletely processed transcripts are difficult to distinguish from regulated processing (Filichkin, Priest et al. 2010). At the library preparation stage, sequence-dependent variation in amplification generates heterogeneous coverage artifacts (Jiang, Schlesinger et al. 2011; Roberts, Trapnell et al. 2011) that makes calling alternative splicing based on exon counts problematic. At the alignment stage, reads with sequencing errors derived from regions that differ in their uniqueness relative to the reference genome (such as paralogs and low sequence complexity regions) confound abundance differences with alignability (Garber, Grabherr et al. 2011). Computational tools to resolve these difficulties are only recently becoming available. In the next chapter, these tools are reviewed, along with a description of a novel tool I have developed.

1.4.3 Methods of RNA-Seq analysis

Tools for analyzing RNA-Seq data are available for several core tasks, such as alignment, gene and transcript expression estimation, and assembly (Garber, Grabherr et al. 2011). Several programs perform the task of spliced alignment well, each with different error profiles (Grant, Farkas et al. 2011). Gene expression estimates for annotated genes also perform well, since they rely on enumerating reads within defined boundaries (Anders and Huber 2010; Trapnell, Williams et al. 2010).

However, there is a dearth of tools that adequately resolve splicing differences detected by RNA-Seq experiments. It is more difficult to quantify splicing is than gene level expression because isoform abundance is often convoluted by overlapping genomic coordinates. An additional problem is that splicing variants are more incompletely annotated than genes, and the true extent of splicing is unknown. Since RNA-Seq error profiles are also incompletely understood, it is challenging to resolve novel detect splicing from technical artifact, particularly when one is analyzing mutants where splicing may be aberrant.

In the next chapter, I describe a tool (Spanki – the Splicing Analysis Toolkit) that addresses these challenges. This tool provides realistic estimates of RNA-Seq error profiles, and allows confident analysis of splicing at the level of individual introns and at splicing events. I also show how it can be applied to successfully resolve sex-differential splicing in wildtype *D. melanogaster*, mutant lines, and multiple Drosophila species; providing novel insight into this model system of transcriptome regulation.

Chapter 2: A Junction-based Splicing Analysis Toolkit (Spanki)

Section 2.1 Abstract

The production of many transcript isoforms from one gene is a major source of transcriptome complexity. RNA-Seq experiments, in which transcripts are converted to cDNA and sequenced, allow the resolution and quantification of alternative transcript isoforms, however, methods to analyze splicing are underdeveloped and errors resulting in incorrect splicing calls occur in every experiment. We demonstrate that these errors include false alignment to minor splice motifs and antisense stands, shifted junction positions, paralog joining, and repeat induced gaps. We developed the Splicing Analysis Kit (Spanki) as a fast, open source, suite of tools. Spanki quantifies splicing differences in transcripts from the same loci within a sample, as well as between samples by using only those reads that directly assay splicing events (junction spanning reads). Despite the reliance on a fraction of the total data, sequencing depth typically generated in an RNA-Seq experiment is sufficient to identify differentially regulated splicing, and error profiles are superior. Critically, Spanki uses simulations based on real RNA-Seq data to characterize errors in a given dataset, and user tunable filters that eliminate those errors. We demonstrate that our computational approach outperforms several commonly used approaches in an analysis of sex-differential splicing in *Drosophila* heads. Spanki can also be used to improve performance of existing tools. The software is available at http://www.cbcb.umd.edu/software/spanki.

Section 2.2 Background

Alternative splicing generates different RNA molecules from identical primary transcripts, affecting protein diversity by creating diverse mRNA isoforms and modulating regulatory information in non-coding and untranslated regions in mRNAs (Black 2003). The advance of next-generation sequencing technologies has allowed the high-throughput analysis of whole transcriptomes by RNA-Seq. In a typical RNA-Seq experiment, Poly-A+ transcripts are enriched from a pool of RNA, from which cDNA is generated, amplified, and sequenced (Oshlack, Robinson et al. 2010). Analysis of RNA-Seq data entails inferring the transcript molecule corresponding to each read, along with estimation of relative abundances of transcribed and processed features (Oshlack, Robinson et al. 2010; Martin and Wang 2011). Thus, we now have the tools to make tremendous progress on understanding mRNA diversity generated by splicing.

Despite the promise, there are important sources of ambiguity, bias, and noise in RNA-Seq data that have made accurate abundance estimation of isoforms difficult in practice. These problems arise at multiple steps in an RNA-Seq experiment. For example, at the biological level, introns retained in incompletely processed transcripts are difficult to distinguish from regulated processing (Filichkin, Priest et al. 2010). At the library preparation stage, sequence-dependent variation in amplification generates heterogeneous coverage artifacts (Jiang, Schlesinger et al. 2011; Roberts, Trapnell et al. 2011) that makes calling alternative splicing based on exon counts problematic. At the alignment stage, reads with sequencing errors derived from regions that differ in their uniqueness relative to the reference genome (such as

paralogs and low sequence complexity regions) confound abundance differences with alignability (Garber, Grabherr et al. 2011). We describe a pipeline and a suite of tools called Splicing analysis kit (Spanki) to make meaningful, comprehensive comparisons of splicing regulation from splice junction reads in RNA-Seq data. These tools provide estimates of sequencing error, metrics
to assess variability and uncertainty in junction detection, classifiers for pairwise splicing events, and generate significance measures of between-sample differences.

A common approach to examining splicing is to determine read coverage of alternative exons, assemble full length isoform models, and generate probabilistic abundance estimates of the alternative forms (Garber, Grabherr et al. 2011). The problem with this type of approach is that reads mapping to exon space may originate from multiple alternative exons with different exon boundaries (Figure 1A). This has been recognized as an inherent problem with short read technology (Oshlack, Robinson et al. 2010). In contrast, reads that span splice junctions derive unambiguously from one exon join, making this a much more useful measurement (Grant, Farkas et al. 2011). However, mapping these reads is more difficult than alignment to a contiguous genomic reference, and high quality junction alignments are critical for downstream analyses that use these alignments (Grant, Farkas et al. 2011). To ensure accurate junction quantifications, Spanki performs simulations to generate sequencing error models, uses novel filtering methods for robust junction detection and quantification, and applies standardized splicing event ontologies to quantify and compare splicing events between samples using high-confidence junction calls.

As a test case for Spanki, we analyzed splicing in Drosophila female and male heads. We chose these samples for two reasons. First, the central nervous system of many species is highly complex in architecture and is a rich source of alternative transcripts (Li, Lee et al. 2007). Additionally, the Drosophila sex determination hierarchy is a classical model of regulated alternative splicing. Three members of this hierarchy, *Sex-lethal* (*Sxl*), *transformer* (*tra*), and *male specific lethal 2* (*msl-2*) are broadly expressed. The two terminal members of the hierarchy *doublesex* (*dsx*) and *fruitless* (*fru*) are expressed in a restricted set of neurons, in addition to other non-neuronal tissues. All of these sex determination genes are regulated by sex-differential alternative splicing (Venables, Tazi et al. 2011). We demonstrate that our approach produces alternative splicing measurements that are consistent with the literature and quantitative PCR (qPCR) results. In benchmarking tests, Spanki provides more precise estimates of pairwise splicing differences than estimates based on transcript level abundances or exon-level counts.

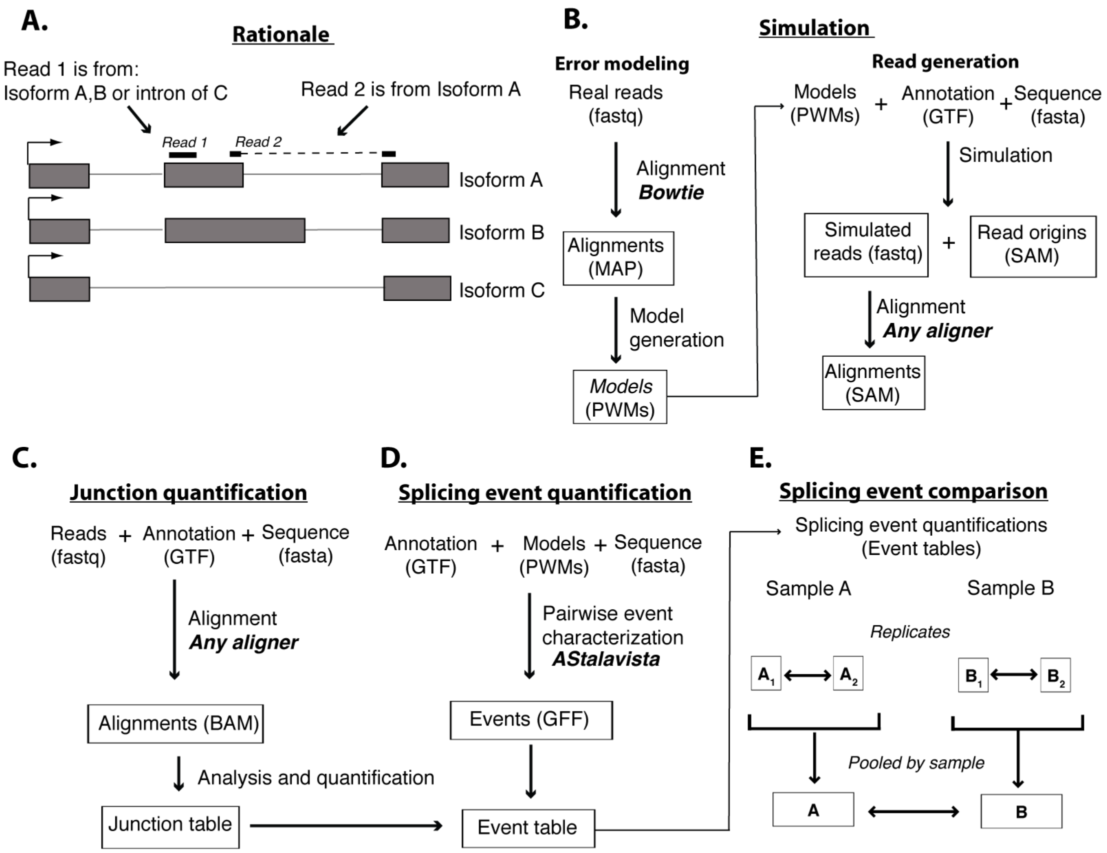


Figure 2.1: Rationale and overview of analysis approach.
Illustration of an analysis using Spanki. (A) Cartoon of a hypothetical locus encoding alternatively spliced transcripts, illustrating how junction-spanning reads map unambiguously to specific introns. Read 1 could have originated from the 2nd exon of isoform A or B, or the intron of isoform C; while read 2 could only have originated from isoform A and the indicated splice junction. (B-E) Flowcharts of analysis steps. For each step, the input data required is listed at the top, with the required format in parentheses. External programs used are indicated in bold. (B) Flowchart of simulation methods. A two step process begins with modeling error profiles based on a permissive Bowtie (Langmead, Trapnell et al. 2009) alignment. These error models are used by the simulator to generate reads. Any aligner can be used to align the simulated reads to a genomic reference, and the aligned positions are compared to known input. (C-E) Flowcharts of quantification and comparison methods. The first step is junction quantification (E), where alignments are performed, junction alignments are curated, and junction coverages are calculated. Splicing event quantification (D), where a set of transcript models (from annotation or computed using a program such as Cufflinks (Trapnell, Williams et al. 2010)), are used to characterize pairwise splicing differences ("splicing events"). These events are merged with junction coverage data to quantify the mutually exclusive paths defined for each event. Splicing event comparison (E) uses these tabulated event-level quantifications to compare between replicates, and between pooled results for each sample, by Fisher's Exact Test on inclusion and exclusion junction counts.

Section 2.3 Results and discussion

To generate the biological data for testing Spanki, we produced pools of poly-A+ RNA from heads of male and female wild-type flies (see Materials and methods for details), and generated two biological replicate libraries for each sex. Sequencing was performed on either GAIIx or HiSeq instruments (Illumina, San Diego, CA) to yield 200 million mapped 76 bp paired-end reads for the female sample, 202 million for the male sample (Table 2.1). We used Spanki to generate error profiles of our data and perform simulations to analyze junction detection performance. These simulations were used to filter false positives and produce a high-confidence set of junction coverage values. We then used Spanki to model, quantify, and classify splicing differences in the RNA-Seq data.

Table 2.1: Mapped reads for Wild type Drosophila heads. All runs are 2x76bp paired-end.

Lane ID	Instrument	Sample ID	Instrument	Total UNIQUE mapping	GEO accession
R50L5_WT_F	GA Iix	WT_F_rep1a	GA IIx	29,637,147	GSM928376
R57L4_WT_F	HiSeq 2000	WT_F_rep2a	HiSeq 2000	62,233,181	GSM928383
R63L5_WT_F	HiSeq 2000	WT_F_rep1b	HiSeq 2000	108,172,410	GSM928392
Total				**200,042,738**	
R50L6_WT_M	GA Iix	WT_M_rep1a	GA IIx	29,484,201	GSM928377
R57L5_WT_M	HiSeq 2000	WT_M_rep2a	HiSeq 2000	61,541,082	GSM928384
R63L6_WT_M	HiSeq 2000	WT_M_rep1b	HiSeq 2000	111,146,748	GSM928393
Total				**202,172,031**	

2.3.1 Analysis of junction detection

Since junction detection is the foundation of our analysis, we undertook simulations to quantitatively assess splice junction detection performance so that we could characterize and then filter out dubious junctions. We built simulated datasets in two steps: modeling and read generation. Spanki automates each of these steps to allow the generation of custom simulations to approximate individual RNA-Seq runs (Figure 2.1B).

In the first step, reads from our RNA-Seq experiments were aligned to the genome. We did a first pass alignment with permissive parameters (quality aware alignment, with no fixed mismatch cutoff) using Bowtie (Langmead, Trapnell et al. 2009) in order to estimate total mismatch profiles along the full length of the reads. As has been previously reported, we observed increased mismatch rates extending through the 3 prime end of the read and a slight increase in mismatch rates in the first 5 bases of the reads (Mortazavi, Williams et al. 2008; Li, Ruotti et al. 2010; Jiang, Schlesinger et al. 2011). This pattern was consistent with each replicate of our head data (Figure 2.2A). We also determined frequencies of each nucleotide mismatch to generate a non-random substitution matrix.

In the second step, we supplied these error models to Spanki's read simulator so that we could detect errors in a defined known input sample generated in silico from annotated transcript models (Ensembl release 67, May 2012; corresponding to Flybase 5.39). We extracted transcript sequence from each of these models and generated pools of simulated reads, with each pool containing reads from every transcript at the same coverage, using 76 bp paired-end reads to mirror our real data.

Thirteen pools were generated, at coverages from 1-30X. To model retained introns due to either regulation or incomplete processing, we generated 20% of the reads from each transcript model with introns included. This is an elevated rate of intron retention (empirical estimate is 6.9 - 7.2%, This study), intentionally applied to increase aligner error. For each read pair generated, a fragment size was randomly selected from a normal distribution of mean 200bp and standard deviation 20bp, and a random start position was selected. Modeled error frequencies were applied as weights for mismatch number, position, and substitution. To enable the tracking of aligner errors, we incorporated the genomic coordinates of origin for each read into a unique read identifier. We calculated two consensus quality scores across all positions from the empirical data - one for matched positions, and one for mismatched positions, and used these to generate a quality string for each read. These components were merged together to output reads in FASTQ format, along with a SAM file that represents a perfect alignment of the reads. We then uniquely aligned the reads using TopHat (Trapnell, Pachter et al. 2009), and compared alignment results to the known input to explore splice site detection parameters. This two-step process generated a simulated data set that mirrored our experimental data and where the true input was known, which provided us a platform for testing RNA-seq junction alignment. Both of these steps (error modeling and read generation) are integrated within Spanki so that we could evaluate detection and quantify differences using the same set of tools (Figure 2.1B-E).

2.3.2 Detection sensitivity and accuracy

Since Spanki is dependent on junction coverage to estimate splicing event abundances, we evaluated the quantification accuracy of detected junctions in the simulations. We compared junction coverage detected by TopHat with known input abundance for all junctions in the 10x transcript coverage pool (Figure 2.2B). Since multiple transcripts at a locus may share a given junction, individual junction coverage spanned a range of 1x to 400x (median 8x, 4.2 million read pairs) reflecting both the random sampling of read positions and overlapping Drosophila transcript models at a given locus. Our junction coverage measurements had high concordance with simulated input (Pearson's r = 0.99, Figure 2.2B) demonstrating that junction coverage closely tracks known input. A lone outlier in this concordance were transcripts from the gene *para*, a complex locus with known RNA editing (Hanrahan, Palladino et al. 2000).

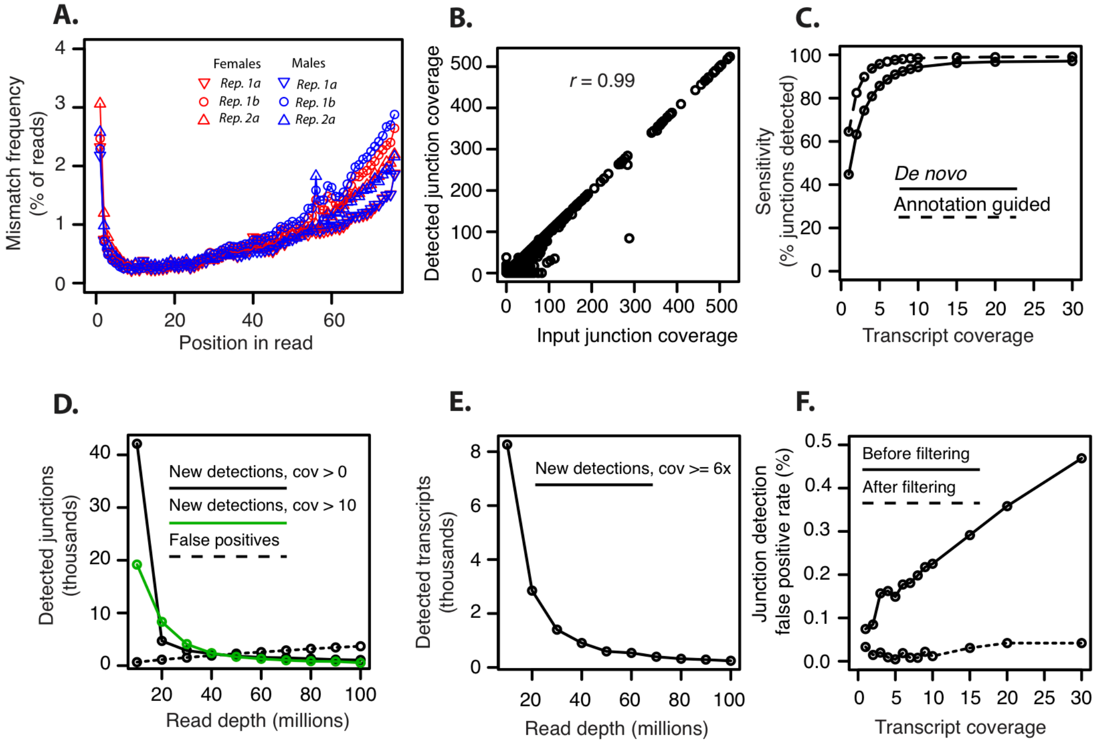


Figure 2.2. Simulation results and junction detection.
Evaluation of junction detection by simulation and by subsampling of real data. (A) Mismatch frequency by position in read in real data. Results for each replicate (technical and biological) of female samples (red lines and symbols) and male samples (blue lines and symbols) are indicated. (B) Accuracy of detection at annotated junctions. Recovered junction coverage after mapping simulated reads (y-axis) is compared to actual coverage in simulated input (x-axis). (C) Sensitivity of junction detection. Receiver operator characteristic (ROC) curve of splice junction detection displays sensitivity of junction detection as it relates to sequencing depth. Results represent TopHat mapping with a supplied annotation ("Annotation guided", dashed line), and without an annotation ("*De novo*", solid line). (D) Junction detection in subsamples of real data. Junction detection in read pools of increasing sequencing depth (10-100 million reads in increments of 10 million). Junctions detected in each pool with at least one read (black line), or robustly detected with 10 reads (green line) are indicated. For each pool, the additional junctions detected relative to the previous pool are indicated. Total false positive junction detections in each pool (dashed line) are also plotted. (E) Transcript coverage in subsamples of real data. New transcripts detected with at least 6x coverage (black line) in each subsampled pool of real data is plotted. (F) Junction detection false positive rate in simulated data before filtering by Spanki (solid line) and after filtering (dashed line).

Junction spanning reads are a small portion of the total reads in an RNA-Seq experiment (9.4 - 12.6% in the six samples used in this study) raising the possibility that sufficient coverage for calling junctions would be problematic. To test for the effects of read depth, we generated pools of simulated reads for each annotated reference transcript at multiple fixed coverages (1-10x, 15x, 20x, and 30x) and aligned these simulated reads with ('Annotation guided') or without ('Denovo') a reference annotation, and compared detection results with known input (Figure 2.2C). We detected >90% of junctions with 3x simulated transcript coverage when we provided an annotation to the TopHat aligner. Without the benefit of annotation, we found that 6x coverage was required to reach this level of sensitivity. Reaching this level of coverage for each annotated transcript (63 million bp of transcript sequence) required 2.5 million read pairs (5 million total reads). For each sample of our experimental data, we obtained at least 200 million total reads. However, as we explain later, real biological samples contain transcripts in unequal proportions, so obtaining high coverage of a rare transcript is difficult. To put this in context of our experimental data, we detected 8,266 transcripts at coverage >= 6x with 5 million mapped reads.

To relate our results on sensitivity to real data, we simulated different sequencing depths by sampling in 10 million read increments from one high-depth experiment (female heads, Sample ID: WT_F_rep1a, Table 2.1) by random selection (without replacement), and evaluated junction detection in each pool. We found that > 40,000 junctions (> 65%) were detected in the first 10 million reads and that a 10-fold greater read depth added ~20,000 more junctions (91% of the total junctions detected at 200 million reads) (Figure 2.2D). At depths of > 50 million mapped reads, the number of false positive detections exceeded the number of new junction detections (Figure 2.2D), as well as the number of new junctions detected robustly (>10 reads). This shows that when we exceed this depth, we begin to detect more false positives than new true positives. Also at this depth, the number of new transcripts detected with at least 6x coverage begins to level off (Figure 2.2E), and we obtain 6x coverage of 95% of the transcripts reliably detected at FPKM >= 1 in the full dataset (200 million mapped reads). These data indicate that simply increasing read depth in a sample results in rapidly diminishing returns of detected splice junctions. This has obvious implications for experimental design and sequencing strategy.

2.3.3 False positive junctions and filters

Since every genome is incompletely annotated, RNA-Seq experiments are likely to reveal splice junctions that are not yet annotated. Distinguishing novel detection from experimental error is a major challenge of RNA-Seq analysis. In our simulations of annotated transcripts, any unannotated junction detected was a false positive, which allowed us to estimate junction detection false positive rates in real data. We examined the false positive rate at multiple transcript coverages (Figure 2.2F) and found that the rate increased with greater transcript coverage due to cumulative errors in alignment. These data indicate that greater read depth provides more opportunities to call false positives in addition to the diminishing returns outlined above. Even though the false positive rate was < 0.5% of all detected junctions (up to 30x transcript coverage), with tens of thousands of junctions detected in an RNA-Seq experiment, even these low error rates generated hundreds of false positives. Junction detection errors have far-reaching downstream effects such as calls incorrectly supporting gene merges, antisense transcripts, and alternative splicing events.

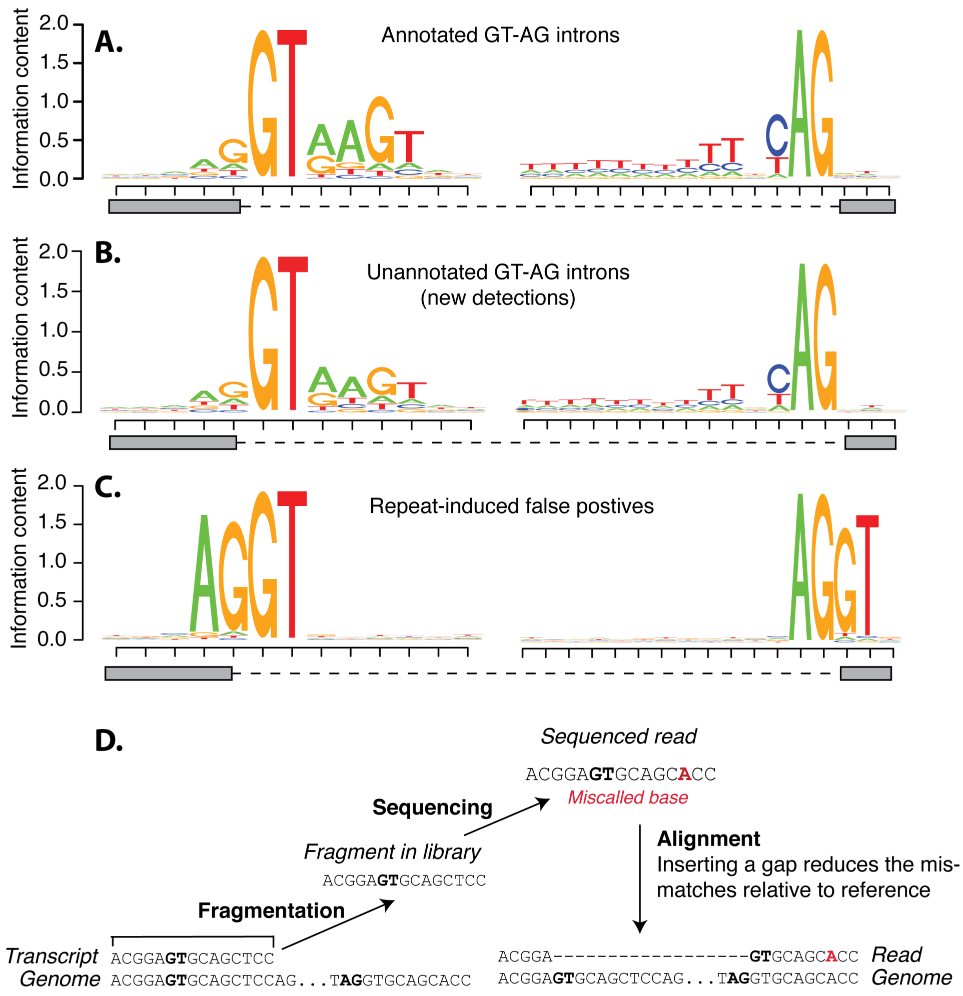


Figure 2.3: Sequence characteristics of false positive junctions.
Sequence logos of exon and intron sequence bordering splice junctions in (A) annotated GT-AG introns, (B) unannotated GT-AG introns detected that pass filtering, and (C) repeat induced false positives. (D) Cartoon illustrating how false positives arise from repetitive sequence and sequencing error. A transcribed fragment from a region of repetitive sequence is incorporated into a library. A base calling error (in red) produces a read with an “A” instead of a “T” at the indicated position. This incorrect base call induces an incorrect gapped alignment that minimizes sequence mismatches.

To lower the false positive rate, it is important to understand the nature of the errors. We therefore examined the sources of alignment error that lead to false positives at 30x coverage (Table 2.2). The dominant source of error was due to the aligner using minor acceptor donor motifs rather than the canonical motif because mismatch reduction takes precedence over the relative likelihood of motifs. The most common donor / acceptor motif pattern is GT-AG, and these major forms have additional well-defined motifs within the intron sequence (Figure 2.3A). However, minor forms have been described (Hall and Padgett 1996) and two of these (GC-AG and AT-AC) are detected by TopHat (Trapnell, Pachter et al. 2009). Although AT-AC introns are > 100X rarer than GT-AG introns in the annotation, TopHat chose the more optimal alignment, resulting in the false placement of a GT-AG spliced alignment on a proximal AT-AC site because of an alignment with fewer mismatches at a proximal AT-AC site than to the correct GT-AG site.

The preference for optimal alignment with fewer mismatches also led to false positive alignments on incorrect strands. In RNA-Seq data from non-strand-specific protocols, the strand is inferred from the sequence of the interior donor/acceptor motif. We observed 78 cases where an incorrect alignment occurred on the opposite strand of the simulated transcript sequence (31.6% of false positives, Table 2.1). For example, a shift in the 3 prime end of the alignment can cause a (+) strand GT-AG intron to be read as a (-) strand minor form GT-AT intron. If uncorrected, errors of this type can lead to the false prediction of antisense transcripts.

Table 2.2: Sources of false positive junction detection

Type of error	False positives	Qualitative filtering strategy	Removed by qualitative filtering[1]	Removed by quantitative filtering[2]
False alignment to minor form	36.40%	Remove novel minor forms	36.40%	30.70%
Incorrect strand	31.60%	Inconsistency with gene model	31.60%	28.80%
Shifted on same strand	13.80%	None	0%	12.20%
Paralog joining	8.50%	Inconsistency with gene model	8.50%	7.70%
Repeat sequence induced	7.70%	Exon-intron sequence similarity	7.70%	6.50%
Unidentifie d error	2%	None	0%	0%
Total defined errors:	100%	***Total removed errors:***	84.20%	75.90%

[1] False positives removed by Spanki's qualitative filtering
[2] False positives removed by filtering on entropy score (>= 2), calculated by Spanki

A subtler error type we observed was the placement of a junction alignment shifted from its correct location on the same strand (13.8% of false positives, Table 2.1). In these error types, mismatches induce a misplaced alignment over a major form GT-AG intron that is consistent with the strand of the simulated transcript. Within this class of errors, 33% of them correctly place at least one end of the alignment (the donor or the acceptor) in the correct place, and 12% of them incorrectly join annotated donors and acceptors from different transcripts of the same gene.

Another error class we detected was the joining together of exons in paralogs as if they came from one gene, rather than keeping these as distinct transcripts from different genes. Paralogs often reside proximally in the genome and retain a high degree of sequence similarity. This similarity led to errors, where a splice junction originating from one paralog was aligned as a join between separate paralogous genes. Although this is a smaller class of errors (8.5% of false positives, Table 2.2), they falsely suggest the presence of merges of distinct genes into a single gene model.

While all the preceding error types resulted from incorrect placement of spliced alignments, we also observed cases where spliced alignments were incorrectly induced in reads that originated contiguously from the genome, resulting in the inappropriate insertion of an intron into an input exon (7.7% of false positives, Table 2.2). This error type occurred most frequently (78% of occurrences) in low sequence complexity regions with either very high or very low GC% (>70%, or <10%). In these regions, mismatches induced a more optimal alignment when the read was split and joined to another segment up or downstream. This type of error can be clearly seen when comparing the extended donor/acceptor sequence motifs of these false positives to annotated introns. False positives occurred in repetitive sequence, where the incorrect 5 prime intron sequence was highly similar to the anchor sequence in the 3 prime exon. In this type of error, we saw an over representation of the motif "GTAG" on both ends of the junction (Figure 2.3C,2.3D). These are a minor class of false positive in simulated transcript sequence (7.7% of false positives, Table 2.2).

We performed an additional simulation consisting only of contiguous genomic sequence, to estimate the frequency of this class of error should there be contaminating intergenic sequence in a sample. 10 million simulated reads from contiguous genome sequence resulted in 310 false positive junction alignments. Thus, in an RNA-Seq experiment with contamination from genomic DNA, or robust intergenic transcription without splicing, repeat-induced errors will be generated at a rate of 1 per 36,000 intergenic-derived read. These errors create the appearance of introns in intergenic noncoding RNAs, and given that an intron is often used as evidence for a transcript and not contaminating DNA, these errors can lead to false calls of intergenic transcription.

After characterizing the above sources of error, we sought to filter and remove as many as reasonably achievable. We built several filtering criteria into Spanki that address the specific error types described above, and examined their effectiveness at removing errors.

We first examined the effectiveness of a simple quantitative cutoff on the alignment entropy score (Graveley, Brooks et al. 2011). This metric quantifies alignment complexity based on diversity of alignment offsets. Requiring a minimum entropy score of two for each junction removed 75.9% of the false positives we identified (Table 2.2). However, since quantitative filtering criteria may be overly stringent in the case of rare transcripts, we developed qualitative criteria that allow filtering of low abundance junctions.

To prevent strand switches and gene merges at paralogs, we generated gene “assignments” for each end of a splice junction, by finding genes that overlap the same region and strand as the putative junction. An overall gene assignment was made from the results for each end of the junction spanning read, and junctions were flagged as "ambiguous" and filtered out if each edge was assigned to a different gene or if either end was assigned to no gene. We found that filtering on this simple criterion was effective in removing all false positive junction detections in simulated data where a junction was called on the wrong strand or if paralogs were incorrectly joined (40.1% of false positives, Table 2.2). To filter repeat sequence induced errors, we used the edit distance between exon shoulder sequence and intron sequence. For each junction, Spanki compared 10bp upstream of the donor to 10bp upstream of the acceptor, and 10bp downstream of the donor to 10bp downstream of the acceptor, and reported the percent identity. Using a threshold of 80%, this comparison revealed cases where similarity between putative exon and intron sequence generated false gapped alignments. We found that filtering junctions where introns were > 80% identical to up or downstream exon sequence removed these errors (7.7% of false positives, Table 2.2). To remove cases where mismatches induced alignment to a minor form intron, we removed introns of this minor class when they were not annotated (36.4% of false positives, Table 2.2).

Applying the qualitative filtering criteria above removed 84.2% of false positive junctions in our simulated data (Table 2.2). This removed 8.4% more false positives than using entropy scores alone, without requiring junctions to be detected with high coverage. This led to an overall false positive rate of < 0.04% across all simulated read depths when using Spanki. We next applied these filters to our experimental data, to define a set of high-confidence junction detections in female and male heads. All of these filters are user tunable and can be adapted based on the experiment.

2.3.4 Splicing detection in D. melanogaster heads

We assayed splicing in *D. melanogaster* heads by analyzing splice junctions detected in our RNA-Seq read alignments. Spanki quantified 70,827 junctions in heads passing our false positive filtering criteria, of which 24,711 were unannotated in Flybase r5.39. We examined the sequence motifs of the major form (GT-AG) unannotated junctions detected in heads, and found them to be nearly indistinguishable from the motifs of annotated junctions (Figure 2.3A,2.3B), with clear branch sites and polypyrimidine tracts. Using Spanki's gene assignments for these junctions, we found that they arose from 5,329 genes, and no single gene contained more than 1% of the total unannotated junctions, showing that novel junction detection was not due to under-annotation of a small group of genes. These data suggest that transcript diversity is under-annotated.

Genes with low abundance transcripts pose a problem for downstream differential analysis, as coverage is reduced due to both splicing differences and primary transcript abundance. To determine whether low coverage junctions are due to rare splicing or low levels of transcription, we examined the junction-level "inclusion rate" calculated by Spanki (see Details of software design, Figure 2.4A).

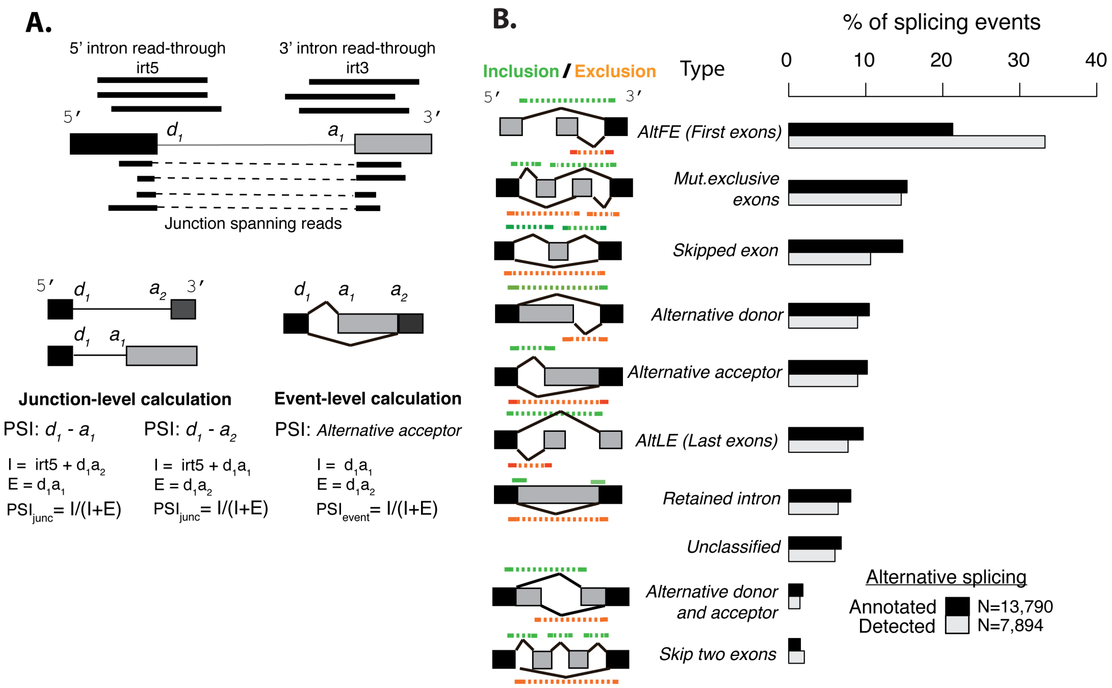
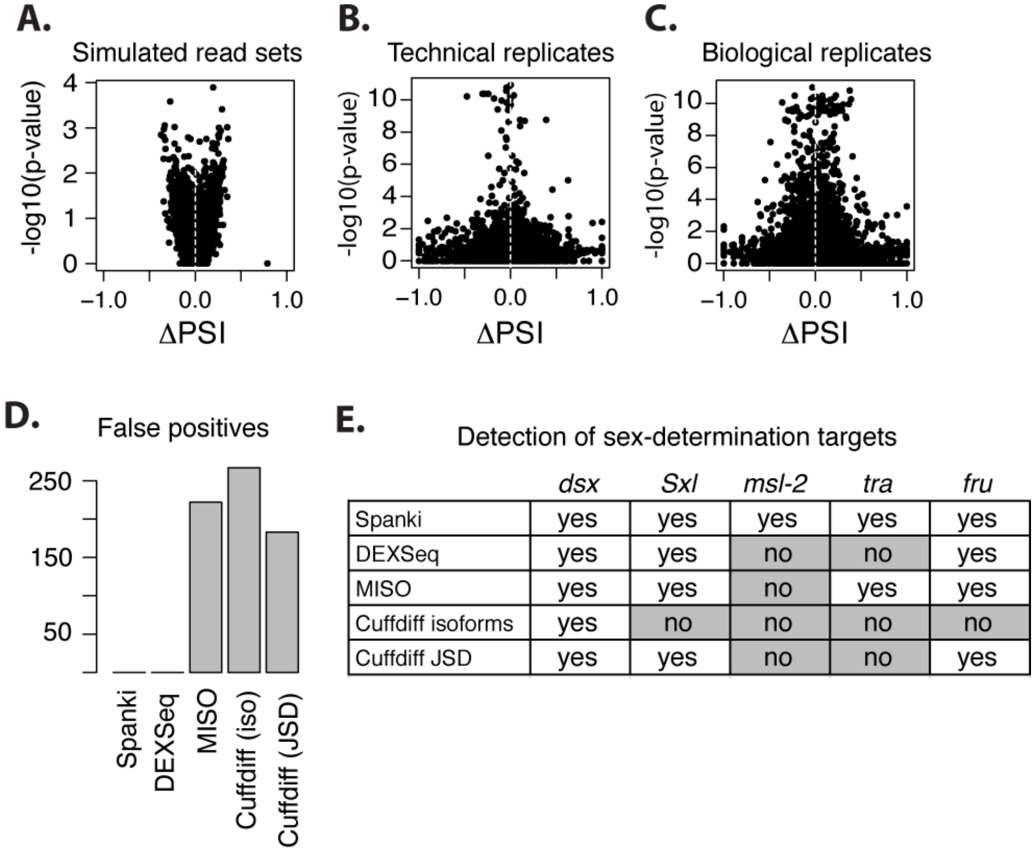


Figure 2.4: Calculation details and splicing event characterization.
Details of how Proportion Spliced In (PSI) is calculated, and results for each splicing event type. (A) An example pair of splice junctions, where a donor is shared between two alternative acceptors. For each junction, intron read-through is calculated for the five prime (*irt5*) and three prime (*irt3*) ends. A proportion spliced in (PSI) is calculated for each junction (PSI_{junc}), where exclusion is the number of junction spanning reads, and inclusion is the sum of the irt5 value and the number of reads spanning the alternative junction from the same donor. An alternative acceptor "event" is defined that composes the two junctions, and a PSI is calculated for the event ($\text{PSI}_{\text{event}}$) where the number of junction spanning reads over each join is used to define inclusion or exclusion. In each formula, "I" represents inclusion and "E" represents exclusion. (B) Pairwise splicing events defined for all transcript models in Flybase 5.39 annotation (black bars), and the subset of those detected as alternatively spliced in female and male *Drosophila* heads by Spanki (grey bars). Black bars indicate splicing event classes as a percent of all defined pairwise events. Grey bars indicate splicing events detected as a percent of all detected events. Cartoons of each event type are in the leftmost column, with the "inclusion" form indicated in green, and the "exclusion" form indicated in orange. A description of each type is adjacent to each cartoon. The "Unclassified" type includes diverse complex type with no concise verbal description.

This value summarizes coverage at each donor to estimate an inclusion rate for each junction, to identify those that are rare (approaching 0%) or common (approaching 100%). We found that for annotated splice junctions, median inclusion rate was high in females (89%) and males (90%), while for unannotated junctions, median inclusion rate was low (1% in females, and 0.5% in males), clearly demonstrating that the unannotated junctions were rare events. Since junction detection in only one sex may indicate regulated splicing, we compared junction detection in both sexes. We found that 12.5% of junctions were detected only in females, and 14% were detected only in males. However, this does not necessarily mean that sex-biased splicing occurred at this rate, as sex-specific junction detection may result from differences in sampling error, sequencing depth, or transcription, not splicing. Additionally, biological variance was high for many of the sex-specific junctions in this first-pass analysis. We discuss this further when we directly compare splicing in female and male heads.

Before comparing female to male heads, we needed to classify events into alternative exon sets. These pairwise classifications can be defined as: cassette exons, mutually exclusive exons, alternative donors, alternative acceptors, alternative first and last exons, and retained introns (Black 2003). However, there are other classes of splicing events that do not fit into these categories. To analyze the full repertoire of splicing complexity, we expanded these seven basic types by a systematic categorization of all pairwise relationships using the AStalavista tool (Sammeth, Foissac et al. 2008), which constructs graphs from transcript models and outputs complete and non-redundant sets of splicing differences identified through graph alignment (see also Details of software design). This analysis yielded 13,790 pairwise-defined alternative splicing events (Figure 4B). Of these, 9,201 were internal events (not involving the first or last exons) and the remaining 4,589 were alternative promoter events. While alternative promoter use is not alternative splicing per se we include these in our analysis, since isoforms from alternative promoters can be compared when junctions differentiate them. 1306 internal events (14%) did not fit into the seven basic categories. Of this class, “Skip two exons" (200 events) was the largest category, followed by "Alternative donor and acceptor" (two variants, 142 and 137 events, respectively). An additional 827 events (125 unique structures) are termed "Unclassified" because they cannot easily be described in words. The top five occurring structures in the “Unclassified” category comprise 41.5% of these events, each of which represent a variant of a skipped exon event.

Spanki parsed AStalavista output to obtain sets of junctions that define mutually exclusive “paths” (Inclusion and Exclusion, see Details of software design), to identify junctions that interrogate each path specifically. We then used Spanki (Figure 1D) to merge junction coverage data and estimate the relative abundance of the alternative forms. Spanki adopts the Percent Spliced In (PSI) metric (Wang, Sandberg et al. 2008) to express this quantitatively, but reports this as a proportion (ranging from zero to one). Hence, we refer to it here as the Proportion Spliced In (PSI). To find the number of genes alternatively spliced, we selected events for which junction coverage was detected over the inclusion path in either sample, and over the exclusion path in either sample. By this criterion, we found that 7,894 events in 2,441 genes were alternatively spliced in head samples (5,450 internal events in 1,852 genes) (Figure 2.4B). To find events that were sex-differentially spliced, we used Spanki to sum junction data for each inclusion and exclusion path across all replicates to calculate event-level PSI (Figure 2.4A), and performed Fishers exact tests. After correcting for multiple testing, we found 172 events with significant differences between female and male heads (adjusted p-value < 0.01, Benjamini and Hochberg).



	dsx	*Sxl*	*msl-2*	*tra*	*fru*
Spanki	yes	yes	yes	yes	yes
DEXSeq	yes	yes	no	no	yes
MISO	yes	yes	no	yes	yes
Cuffdiff isoforms	yes	no	no	no	no
Cuffdiff JSD	yes	yes	no	no	yes

Figure 2.5: Sources of variation in Proportion Spliced-In (PSI).
Volcano plots where ΔPSI is plotted against the –log10 p-value of the Fisher's Exact Test, to assay variation due to sampling and sequencing error in simulations (A), technical RNA-Seq replicates (B), and biological replicates (C). (A) Plot comparing two simulated read sets of equal reads per kilobase (400 RPK). Since transcript abundances are equal, expected ΔPSI is zero. (B) Variation in ΔPSI between replicate RNA-Seq runs of the same libraries. (C) Variation in ΔPSI between independent biological samples, each with a distinct RNA-Seq library. Results within each sex were similar, results for female samples shown. (D) False positive differential splicing calls in a null dataset. Cufflinks results are shown for both splicing analysis (Jensen-Shannon Divergence, JSD), and isoform abundance comparison. MISO results shown are based on isoform-centric analysis. (E) Summary of detection of sex-differential splicing in components of the sex-determination pathway in Drosophila heads, using Spanki and four other approaches.

To examine potentially confounding sources of variation that are independent of sex, we generated a null model for splicing differences by simulating two pools of reads with equal transcript coverage in Spanki, and also compared technical and biological replicates of real RNA-Seq data to each other. We found minor variation due to sampling alone and technical replication (Figure 2.5A,2.5B), but biological replication was a much greater source of variability (Figure 2.5C), particularly at low total abundance (< 10 average coverage per site in either path). To conservatively adjust for this, we reduced our query set to only events where average coverage per site was > 10 in each path, and the unadjusted p-value for the between-sexes comparison was less than the unadjusted minimum p-value between biological-replicates. We also set a conservative threshold on the difference in PSI (>= 0.20). This filtering yielded 22 events in 17 genes significantly different between the sexes (Table 2.3). The expected members of the sex-determination cascade were identified following filtering. A diversity of splicing types characterizes the core components: alternative donors and acceptors, skipped exons, retained introns, and alternative last exons (Venables, Tazi et al. 2011). The additional targets we detect have a similar diversity of regulation types, with no over-representation of any particular regulation type. Functional sex-biased splicing events might be associated with genes with roles in behavior, as the two terminal sex-determination genes (*dsx* and *fru*) encode alternatively spliced transcripts. Indeed, several of the sex-biased splicing events occurred in transcripts of genes with roles in sexual behavior (see Conclusions).

To compare Spanki's false positive rate relative to other tools, we compared the number of differential splicing calls made in our simulated null dataset. Spanki called zero events differentially spliced in this dataset (Figure 2.5D). We counted reads that map within exons using the script provided with DEXSeq (Anders, Reyes et al. 2012), and performed an exon-level differential analysis. DEXSeq also called zero exons differentially expressed. Next we performed an isoform-centric analysis using MISO (Katz, Wang et al. 2010), which called differential splicing in transcripts of 222 genes (Bayes factor cutoff > 40). Analysis with Cuffdiff (Trapnell, Williams et al. 2010), with default parameters except for specifying upper quartile normalization, called 183 loci as differentially spliced, and 267 isoforms were called differentially expressed. These results showed that Spanki has a low false positive rate of differential splicing calls relative to most other tools.

Since Spanki accurately detected known sex-differential splicing targets, we asked how other tools performed at the same task. Only Spanki detected differential splicing in each target of the sex-determination pathway (Figure 2.5E). DEXSeq (Anders, Reyes et al. 2012), which relies only on exon-level counts, detected significant differences in *fru*, *Sxl*, and *dsx*, but not in *msl-2* or *tra*. Similarly, an analysis with MISO (Katz, Wang et al. 2010) failed to detect differential splicing in *msl-2* transcripts. For Cuffdiff (Trapnell, Williams et al. 2010), we examined results for the splicing difference test (Jensen-Shannon Divergence metric), and also for isoform abundance differences. Neither of these metrics detected differential spicing in more than three out of five targets. These results clearly show that Spanki is superior in quantifying sex-differential splicing in these pathway components.

Genes are often regulated by feed-forward network motifs (Milo, Shen-Orr et al. 2002; Shen-Orr, Milo et al. 2002; Johnston, Otake et al. 2011), so we asked if genes with sex-biased splicing also showed sex-biased transcription. We examined the gene-level transcriptional characteristics of genes expressed in female and male heads and found modest sex-differential expression relative to whole adults, as previously reported by SAGE analysis (Fujii and Amrein 2002) (Figure 2.6). We tested for differential expression using Cuffdiff (Trapnell, Williams et al. 2010), applying upper quartile normalization, and identified 19 genes with sex-differential expression (FDR adjusted p-value < 0.05). To ensure this sparse differential expression was robust to statistical method, we performed a complementary test with the count-based method DESeq (Anders and Huber 2010). Differential expression was also modest by this approach (51 genes FDR adjusted p-value < 0.05), and included all genes called differentially expressed by Cuffdiff. We found that by either approach, genes with sex-differential splicing did not show sex-biased transcript abundance.

Table 2.3: Genes sex-differentially spliced in *Drosophila* heads

Gene ID	Gene name	Event type	ΔPSI[1]	Adj. p-value[2]	GO annotation
FBgn0004652	*fru*	altdonor	-1	5.87E-08	Male courtship behavior
FBgn0003659	*Sxl*	exonskip	0.974	5.87E-08	Sex determination
FBgn0000504	*dsx*	AltLE	0.939	5.87E-08	Sex determination, male courtship behavior
FBgn0004652	*fru*	exonskip	-0.906	9.90E-08	Male courtship behavior
FBgn0028341	*l(1)G0232*	AltFE	0.802	2.98E-08	Protein tyrosine phosphatase activity
FBgn0086675	*fne*	altdonor	-0.656	6.76E-08	Regulation of RNA metabolism
FBgn0005616	*msl-2*	retintron	0.565	1.38E-03	Dosage compensation
FBgn0259923	*Sep4*	AltFE	-0.524	4.79E-04	GTPase activity
FBgn0259923	*Sep4*	altdonor	-0.469	5.87E-08	
FBgn0053113	*Rtnl1*	AltFE	-0.464	6.39E-08	Inter-male aggressive behavior, olfactory behavior
FBgn0053113	*Rtnl1*	AltFE	-0.444	5.87E-08	
FBgn0053113	*Rtnl1*	AltFE	0.426	5.87E-08	
FBgn0004852	*Ac76E*	exonskip	-0.382	5.87E-08	Intracellular signal transduction
FBgn0086674	*Tango13*	altdonor	0.372	6.61E-08	Sulfotransferase activity
FBgn0003741	*tra*	altacceptor	-0.371	5.87E-08	Sex determination, male courtship behavior
FBgn0260660	*mp*	skip2exons	-0.252	7.93E-03	Motor axon guidance
FBgn0259682	CG42351	exonskip	-0.242	9.21E-08	*none*
FBgn0259214	*PMCA*	mutexcl	0.232	5.87E-08	Calcium transporting ATPase activity
FBgn0259214	*PMCA*	exonskip	-0.229	5.87E-08	
FBgn0037297	CG1116	retintron	0.229	1.39E-03	*none*
FBgn0010482	*l(2)01289*	Unclass.	0.22	2.98E-08	Protein disulfide isomerase activity
FBgn0036194	CG11652	AltFE	0.208	8.26E-03	Phagocytosis

[1] PSI in females – PSI in males. Table is sorted by **Δ**PSI absolute value.

[2] p-value from Fisher's Exact Test, FDR corrected by Benjamini-Hochberg

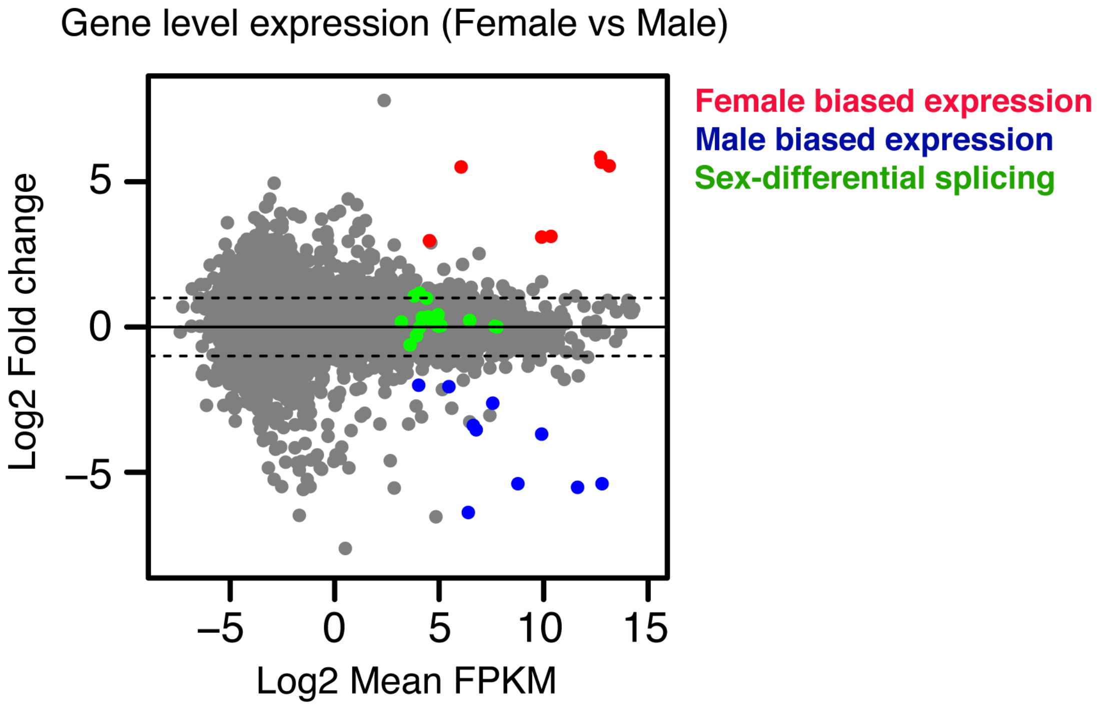


Figure 2.6: Transcription of sex-differentially spiced genes
Ratio vs average abundance scatterplots of gene expression in female and male heads. Log2 fold change (female vs male) is plotted against the log2 mean FPKM (mean of females and males). Genes with significant sex-biased expression (FDR adjusted p < 0.05, Cuffdiff) are shown in red (female-biased) and blue (male biased). Genes with significant sex-differential splicing (**Table 2.3**) are shown in green. No change in expression (solid black line) and two-fold difference (dotted black line) are shown.

Since our splicing event definitions rely on annotated transcript models, we asked whether this restriction prevented the detection of sex-differential events. We analyzed the junction level results from Spanki that test for differences at each donor, independent of a priori knowledge about up or down-stream exon connections (c.f. Figure 2.4A). We extracted unannotated junctions with an adjusted p-value for differential splicing < 0.01, total detected coverage in either sample > 10 reads, and where junction coverage was greater than intron read-through. Through this analysis Spanki found three putative sex-differential unannotated junctions, in *Epidermal growth factor receptor pathway substrate clone 15 (Eps-15)*, *Grunge* (*Gug*), and *bedraggled* (*bdg*). Thus Spanki can also be useful for identifying unannotated events that can then be used to update annotations prior to rerunning the analysis tools.

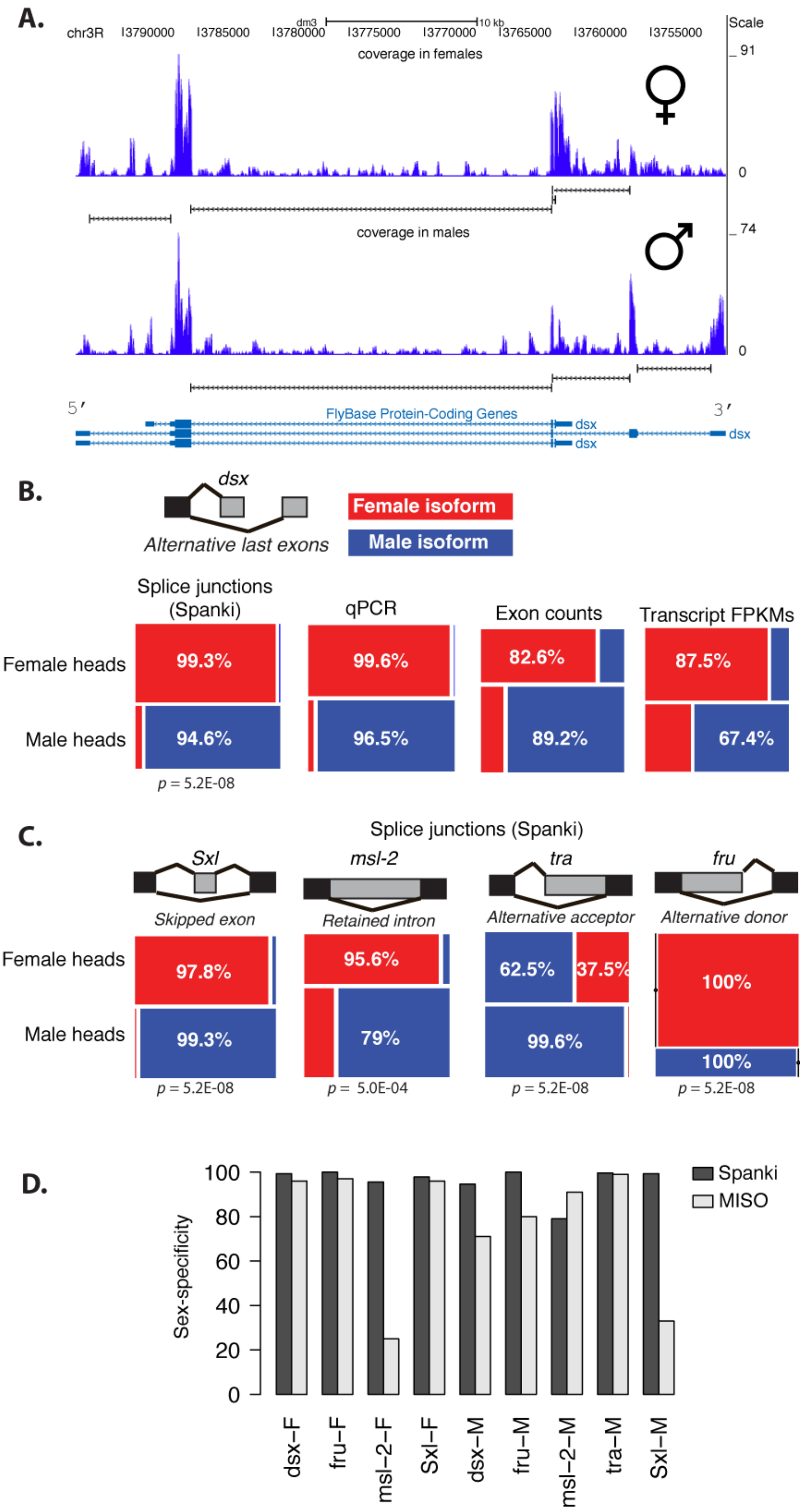
2.3.5 Validation of sex-differential splicing

While our extensive simulations allowed us to tune Spanki using known input, biological samples are known to be unknowns. To test our predictions and stringency, we first compared PSI estimates from Spanki with published estimates for several components of the core sex determination pathway: *dsx*, *fru*, *Sxl*, *tra*, and *msl-2* (Figure 2.7). The transcripts encoded by these genes have been shown to undergo sex-biased (*tra*) or sex-specific (*dsx, fru, Sxl, msl-2*) splicing events (Venables, Tazi et al. 2011). However, this specificity is not clearly visualized in raw base-level coverage results (Figure 2.7A). Spanki's PSI calculation accurately reflected the sex-specificity of these splicing events (calling 79%-100% of the sex-specific isoform) in *dsx*, *fru*, *Sxl*, and *msl-2* (p-value ≤ 5.0E-04) (Figure 2.7B,C). In the case of the sex-biased *tra* splicing event, Spanki detected the presence of the interrupted ORF isoform in females (62.5%) and males (99.6%), as previously observed in Northern blot experiments (Nagoshi, McKeown et al. 1988).

Our splicing calls for the sex determination transcripts are more sex-biased than previous RNA-Seq experiments (Graveley, Brooks et al. 2011) on whole adult flies. To help determine if this was due to methodology, we also quantified splicing events using measurements of exon coverage and isoform abundance estimates (in expected fragments per kilobase of transcript per million mapped reads - FPKMs), to see if these approaches yielded similar results. These metrics predicted results that were much less sex-specific; for example in the case of *dsx* sex-specificity was 67.4 -

89.2% by exon counts or FPKM, and 94.6- 99.3% by Spanki (Figure 2.7B). These results show that using junction coverage with Spanki results in more switch-like splicing difference calls. The MISO tool (Katz, Wang et al. 2010) also produces PSI estimates, so we compared PSI values for all sex-specific splicing events in this pathway between the two tools and found that Spanki showed greater sex-specificity (Figure 2.7D).

Figure 2.7 (Next page): Resolution of splicing differences in the sex determination pathway.
Detection and visualization of sex-differential splicing in sex determination pathway components (Venables, Tazi et al. 2011) by different methods. (A) Genome browser view of alignments within the *dsx* locus, for female (top) and male (bottom) *D. melanogaster* heads. Base level coverage from mapped reads (TopHat (Trapnell, Pachter et al. 2009)) loaded in BED format and visualized in the UCSC genome browser (Kent, Sugnet et al. 2002) (edited for better visibility), for one replicate of data for each sex. Density of reads mapping contiguously to the genome are in blue, and junction spanning reads are shown underneath with brackets. (B) Motif representation of the regulated alternative last exons splicing event in *dsx*, showing the splicing difference between the female isoform (top) and male isoform (bottom), Mosaic plots display the specificity for the female isoform (red) and the male isoform (blue) in each sex, calculated as a proportion from different quantitations of the sex-specific isoforms: splice junctions counts, qPCR, counts of reads within exons, and full length isoform abundance estimates (FPKM). (C) Splicing event motifs for other components of the sex determination pathway, along with their sex-differential splicing results obtained from junction counts with Spanki in mosaic plots: *Sxl*, skipped exon; *msl-2*, retained intron, *tra*, alternative acceptor, and *fru*, alternative donor. Significance measures from Spanki (Benjamini-Hochberg adjusted p-value) are shown beneath each mosaic plot. For each gene, red indicates the female isoform, and blue indicates the male isoform. (D) Bar plot of splicing sex-specificity as quantified by the PSI values report by Spanki and MISO. Results for MISO are from an event-centric analysis.


A.
chr3R
13790000
13785000
dm3
13780000
13775000
13770000
10 kb
13765000
13760000
13755000
Scale
coverage in females
91
0
♀
coverage in males
74
0
♂
5′
3′
FlyBase Protein-Coding Genes
dsx
dsx
dsx
B.
dsx
Alternative last exons
Female isoform
Male isoform
Splice junctions (Spanki)
qPCR
Exon counts
Transcript FPKMs
Female heads
Male heads
99.3%
94.6%
p = 5.2E-08
99.6%
96.5%
82.6%
89.2%
87.5%
67.4%
C.
Splice junctions (Spanki)
Sxl
Skipped exon
msl-2
Retained intron
tra
Alternative acceptor
fru
Alternative donor
Female heads
Male heads
97.8%
99.3%
p = 5.2E-08
95.6%
79%
p = 5.0E-04
62.5%
37.5%
99.6%
p = 5.2E-08
100%
100%
p = 5.2E-08
D.
Sex-specificity
100
80
60
40
20
0
Spanki
MISO
dsx–F
fru–F
msl–2–F
Sxl–F
dsx–M
fru–M
msl–2–M
tra–M
Sxl–M

Increased sex-specificity is only desirable if it truly reflects the biology. For further evaluation of Spanki, we performed quantitative PCR (qPCR) experiments on additional biological replicates, in order to measure the amounts of the inclusion or exclusion forms, respectively, relative to the level of *Actin 5C* transcripts. Within each sex, we compared the inclusion / exclusion ratio in females to the inclusion / exclusion ratio in males. We performed these experiments on *dsx*, *fru*, and ten additional splicing events chosen from among a list of transcripts initially called as sex-biased, but rejected following filtering based on variance in biological replicates and ΔPSI magnitude (Figure 2.8). These events represented each of the basic splicing types, and covered a broad range of PSI values. The two methods were quantitatively highly similar as proportions (Figure 2.8A), and the median value of $\text{ratio}_{\text{qPCR}}$ / $\text{ratio}_{\text{RNA-Seq}}$ was close to one (1.16). These experiments confirmed the high degree of sex-specificity for the *dsx* and *fru* events. The sex-differential splicing in the other transcripts that failed Spanki's filtering criteria showed low magnitude differences in splicing, and sometimes switched direction (e.g. female-biased to male-biased) depending on technique and/or biological replicate (Figure 2.8B). These results also underscore the importance of using filters in addition to statistics for producing robust differential splicing calls.

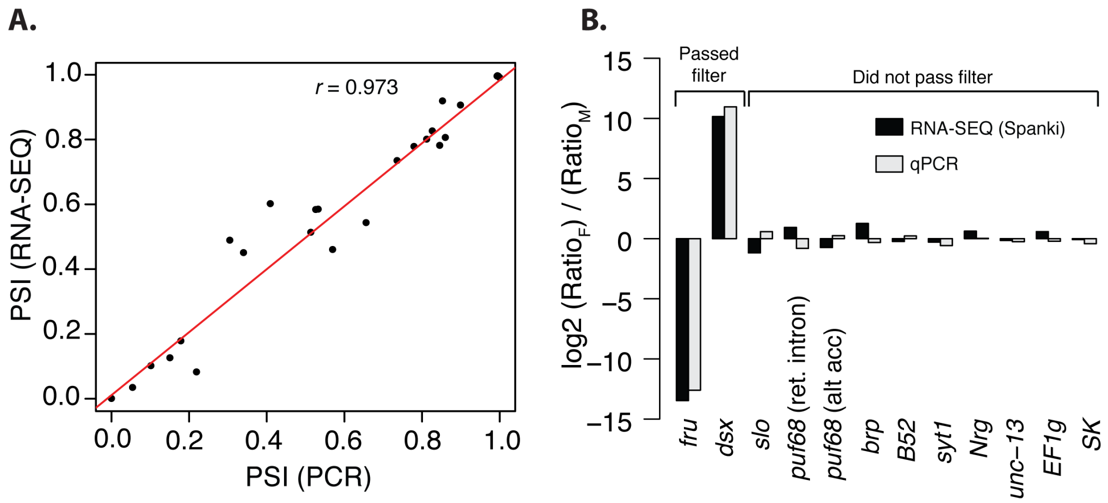


Figure 2.8: Results of qPCR validation.
Comparison of results for alternative splicing quantitation by Spanki (RNA-Seq) and by qPCR. (A) Scatter plot of proportion spliced in (PSI) by RNA-Seq vs qPCR for the inclusion or exclusion form for each event assayed. Inset is Pearson's r coefficient. (B) Barplot of ratio/ratio comparison for RNA-Seq and qPCR. A ratio (of the Inclusion value / Exclusion value) is calculated within each sex. The result is then compared ratiometrically (Ratio$_F$ / Ratio$_M$) and presented for each method (RNA-Seq (Spanki), black) and qPCR (grey), for each splicing event assayed by qPCR. Bars indicate splicing events that passed significance test filtering by PSI magnitude and variance in biological replicates, and events that did not pass filtering.

2.3.6 Comparison to other tools

One strength of Spanki is that is provides diverse functionality in a single toolkit, along with novel quantitative and qualitative junction-level analysis. Simulation tools are available in other programs, and separately, several tools offer differential splicing analysis, using full-length transcript abundances (Cufflinks, (Trapnell, Williams et al. 2010)), read counts within exons (DEXSeq, (Anders, Reyes et al. 2012)) and Bayesian inference from generative models (MISO (Katz, Wang et al. 2010)). The features in Spanki compared to these tools is described in Table 2.4.

Table 2.4: Comparison of features among RNA-Seq analysis tools

Feature	Spanki	Tophat	Cufflinks	MISO	DEXSeq	RUM	Flux Capacitor	Maq
Simulation tools	x					x	x	x
Empirical error modeling	x						x	x
Custom simulated transcript coverages	x							
Junction alignment curation	x	x[1]						
Gene assignment for junctions	x		x[2]					
Qualitative junction analysis	x							
Junction-level comparisons	x							
Event-level comparisons	x		x[3]	x	x[4]			
PSI metric reporting	x			x				

[1] Tophat offers criteria for filtering what is reported after the alignment stage. Spanki provides additional criteria that can be applied after reporting
[2] Cufflinks assembles transcripts and merges with annotated genes
[3] Cuffdiff reports differential splicing by TSS group, without specifiying the differential splicing event
[4] DEXSeq provides results for exon-level abundance differences

To assess Spanki's false positive rate of differential splicing detection, we generated a null dataset using simulated reads and compared differential splicing calls using Spanki and other tools. For this null dataset, we made four read pools, each of which contained reads from all annotated transcripts in equal abundances (300 reads per kilobase of transcript). Each read pool was an independent simulation, where mismatches were introduced into reads using empirical error models. We arbitrarily divided these pools into two groups, replicate 1 and 2 of "Sample A" and replicate 1 and 2 of "Sample B." We then mapped each read pool using Tophat (Trapnell, Pachter et al. 2009), and fed these alignment files to several tools, to test whether calls of differential splicing would result.

Spanki called zero events differentially spliced in this null dataset. Next we performed an analysis with Cuffdiff (Trapnell, Williams et al. 2010), with default parameters except for specifying upper quartile normalization. This program correctly called zero genes differentially expressed. However, 183 loci were called differentially spliced, and 267 isoforms were called differentially expressed. We counted reads that map within exons using the script provided with DEXSeq (Anders, Reyes et al. 2012), and performed an exon-level differential analysis. DEXSeq also called zero exons differentially expressed. These results showed that Spanki and DEXSeq correctly call no false positive splicing differences in the null model.

We have demonstrated Spanki's accuracy with quantifying sex-specificity of known sex-determination pathway components (Figure 2.7), so we next asked how other tools performed at the same task. Only Spanki detected differential splicing in each target of the sex-determination pathway. DEXSeq, which relies only on exon-level counts, detected significant differences in *fru*, *Sxl*, and *dsx*, but not in *msl-2* or *tra*. DEXSeq does not report event-level PSI, so we estimated this metric by using exon-level counts that were normalize by the program. *doublesex* is the only target that demonstrates sex-specificity by this metric, owing to the fact that it is the only event in this list where mutually exclusive exons can be quantified. PSI estimates were calculated for *tra* and *dsx* since there is unique exon space for at least one isoform, but the calculation could not be made for *msl-2* and *Sxl*, which lack unique exon space for either isoform. In the case of *fru*, DEXseq detects differential expression by virtue of higher abundance in the female-specific exon, but reports no difference in the exon space that is shared by both isoforms. For Cuffdiff, we examined results for the splicing difference test (Jensen-Shannon Divergence metric), and also for *inter-se* isoform abundance differences. No significant differences were detected by sex by either of these tests, for any of these sex determination targets. These results clearly show that Spanki is superior in quantifying sex-differential splicing in these pathway components.

Section 2.4 Details of software design

Spanki is an open-source python package distributed under the GNU public license. It is available at http://www.cbcb.umd.edu/software/spanki and all source code can be downloaded from the Github public repository at https://github.com/dsturg/Spanki. It is lightweight and rapid, designed for inter-operability with other open-source tools, and accepts input data in standardized format (BAM, GTF, FASTA), using open-source python modules. It evaluates alignments in BAM format at the rate of 10 minutes per GB, and the remainder of processing time for a typical analysis takes less than five minutes (benchmarked on a 2.8 GHz core i7 iMac with 8GB of RAM).

A complete analysis using Spanki consists of these major steps (Figure 2.1B-E):

- Simulation
- Alignment evaluation and junction quantification
- Generation of junction sets from precomputed splicing event definitions
- Differential testing on junctions and events

2.4.1 Simulation

Performing simulations of RNA-Seq data generation is a common approach to benchmarking tool performance (Trapnell, Williams et al. 2010; Grant, Farkas et al. 2011). Several tools exist to perform simulations with modeled error profiles (BEERS, (Grant, Farkas et al. 2011), maq (Heng Li, http://maq.sourceforge.net/), Flux Simulator (Michael Sammeth, http://flux.sammeth.net/)). The read simulator in Spanki is unique in that it combines robust empirical modeling with detailed reporting that is geared toward evaluating splicing detection performance. This allows the production of simulations that approximate real experimental error profiles, which can be done while a pipeline is under development or for every sample.

<u>Error modeling</u>

Spanki estimates model parameters from a first pass alignment of real RNA-Seq reads using permissive quality aware mapping with Bowtie (Langmead, Trapnell et al. 2009). These alignments allow Spanki to estimate the true error rates within the experimental reads. The error modeling function within Spanki parses the alignments in Bowtie's map format, and produces probability weight matrices for mismatches by position in the read and by base substitution type, and for quality scores by position. The read simulator uses these models to introduce mismatches.

<u>Read generation</u>

Spanki's RNA-Seq simulator function generates simulated reads. The basic input Spanki needs to conduct a simulation is a set of transcripts to simulate, a depth of coverage, and models for incorporating error. Spanki's simulator takes transcript models in GTF format, and extracts transcript sequence from a genomic reference to conduct the simulation. To simulate intron retention, Spanki generates a fraction of simulated reads (specified by the user) from complete transcript sequence where introns are retained. The depth of coverage can be specified by the user in units of coverage or reads-per-kilobase (RPK). For coverage, the number of reads (N) is calculated by the formula N = (C * G) / L, where C is the coverage (eg, 2x), G is the transcript length, and L is the read length. Reads-per-kilobase (RPK) normalizes for feature length so that reads can be generated for transcripts in fixed proportions, creating a null model for splicing differences. Spanki calculates the number of reads to simulate based on the user-specified depth, for each transcript. Alternatively, Spanki accepts a text file where the user can list individual transcripts to simulate at different coverages, which allows simulating fixed quantitative splicing differences between alternative isoforms.

Spanki chooses random positions in transcript sequence to extract reads. Mismatches are then introduced according to the specified model. Pre-built error models are included for the experiments described in this study, a sample from the modENCODE developmental timecourse (Graveley, Brooks et al. 2011) (30-day old whole adult male Drosophila), and a simple weighted-random model. In addition, the user can specify a custom model built on the user's own data. These modeled error frequencies are applied as weights for mismatch number, position, and substitution (Figure 2.2B). Weight matrices of quality scores are used to create a consensus quality values across all positions - one for matched positions, and one for mismatched positions, which are concatenated to create a quality string for the read.

In addition to simulated reads, Spanki reports information that facilitates analysis of alignment and detection (Figure 2.2B). Coverage generated by the simulation for each splice junction is reported, along with read counts for each transcript. To enable the tracking of aligner errors, the genomic coordinates of origin for each read are incorporated into a unique read identifier. The true origin of simulated reads is also reported in a SAM file that represents a perfect alignment, which can be fed to an assembler such as Cufflinks (Trapnell, Williams et al. 2010) to allow the evaluation of error in transcript abundance estimates due to assembly separately from errors in alignment.

2.4.2 Alignment evaluation and junction quantification

Some short-read aligners offer filtering criteria, which are either applied at the alignment stage or the reporting stage. For maximum flexibility, Spanki decouples the alignment and filtering steps, with a tool that applies post-hoc analyses of alignment files. This allows alignments to be performed on multiple data sets generally, with consistent filtering applied later, and allows changing the filtering criteria without re-aligning. Spanki streams through an alignment file produced by any aligner (in standard BAM format), using the Pysam module (Andreas Hager, http://code.google.com/p/pysam/) and calculates junction coverage along with alignment diagnostic measurements. These measurements include the number of alignment offsets, alignment entropy (Graveley, Brooks et al. 2011), and Minimum Match on Either Side (MMES, (Wang, Xi et al. 2010)).

In addition to alignment diagnostic values, Spanki generates calculations that are informative of splicing regulation. For example, intron retention is estimated.

Each junction may exhibit subtle intron retention properties in different biological contexts. For this reason, Spanki counts "intron read-through" reads (Wang, Sandberg et al. 2008; Brooks, Yang et al. 2011) for each junction, regardless of the presence of an annotated retained intron isoform (Figure 2.4A). These are read alignments that span the exon/intron boundary without gaps on either side. To ensure comparability, Spanki enforces an overhang requirement, which is user-tunable, and is applied to both intron read-through and junction calling. These intron read-though values are also used to generate a proportion value for each junction that normalizes for differences in transcription and sequencing depth. For each junction, Spanki sums the intron read-through at the donor with junction coverage to all alternative acceptors as an "inclusion" value, to calculate the Proportion Spliced In (PSI) metric (Wang et al., 2008, Venables et al., 2009, Brooks et al., 2011; Figure 2.4A). This provides an estimate of the parent transcript abundance in the sample where the donor/acceptor pair was available to be joined, which Spanki terms the "inclusion rate", calculated as 1 - PSI_{junc} .

Generation of junction sets from precomputed splicing event definitions Spanki provides utilities for parsing splicing event definitions produced by AStalavista (Sammeth, Foissac et al. 2008), and can be adapted to accept event definitions from other sources such as Ensembl (Koscielny, Le Texier et al. 2009) and Sircah (Harrington and Bork 2008). The AStalavista algorithm begins by decomposing transcript models into "sites," which are exon boundaries. Graphs are built for each gene, where sites are nodes and edges connect them. Edges may therefore correspond to introns or exons. Splicing events are found by identifying subgraphs that have identical sites on the ends, but no common interior sites. This process finds regions of the parent transcript where the donor / acceptor sites of two alternatives are present on a parent transcript, but utilized mutually exclusively in processed transcripts. Spanki uses these event definitions to build mutually exclusive "paths" composed of junctions that interrogate each event specifically. Spanki will also flag and report splicing events that cannot be assayed using only junctions (for example, alternative promoters where there is no differentiating junction).

The terminology of ‘inclusion’ and ‘exclusion’ has been used in the literature to refer to differences in size of the spliced product. Product length does not matter for the purposes of our analysis, so Spanki calls the variant with the most 5 prime differentiating site to be the ‘inclusion’ isoform. This improves consistency, guarantees that all events of the same type will have the same inclusion/exclusion structure, and avoids confusion in cases where alternatively spliced products are of equal length. An exception is made for retained introns, where the retention event is always called the inclusion path.

Spanki coverage from joins to exons that are outside of the event being considered. This is because many gene models are complex, and splicing events cannot always be assayed independently. For events with multiple exons in the inclusion or exclusion paths (such as skipped exons), there may be up- and downstream connections to other exons that could confound the results. To adjust for this, Spanki calculates and reports the junction coverage for first-order neighbors of all interior exons that extend to exons outside the splicing event. This coverage may lead to over-or under-counting of inclusion or exclusion joins within the splicing event. Since our model focuses on discrete and specific measurements, we use this information to indicate the presence of potentially confounding coverage for each event.

Since splicing analysis is a comparison of two alternative events, it is convenient to compare using proportions. The PSI metric that Spanki uses to express proportions has been applied elsewhere to splicing microarrays and RNA-Seq (Wang, Sandberg et al. 2008; Venables, Klinck et al. 2009; Brooks, Yang et al. 2011). In RNA-Seq, the calculated read counts are then divided by the number of 'sites' or positions in each path, to normalize each side of the ratio (Wang, Sandberg et al. 2008; Brooks, Yang et al. 2011). Using only junctions yields more consistent comparisons between events than including exon reads, since the number of positions is constant for events of the same type.

2.4.3 Differential testing on junctions and events

Assessing the significance of differences between samples requires accounting for differences in transcription and sequencing depth. The Fisher's Exact Test (FET) is well suited to this task, since testing proportions accounts for differences in sample totals due to depth or transcription. Spanki constructs 2 x 2 contingency tables from junction counts for each splicing event, to test the null hypothesis that the two samples have equal inclusion/exclusion proportions. The test is performed using the fisher python package v.0.1.4 (Brent Pederson, http://pypi.python.org/pypi/fisher/). FDR correction is performed by the Benjamini-Hochberg method implemented in the StatsModels package (Skipper Seabold, Josef Perktold, http://statsmodels.sourceforge.net/).

To help visualize splicing differences, Spanki includes R scripts to produce mosaic plots, where the relative size of each cell is proportional to real (non-normalized) cell counts (Figure 2.7). Code is also included to produce fourfold plots, which provide a visual test of the null hypothesis of the FET. This provides an effective simultaneous visualization of normalized proportions and significance. These plots are implemented in the "vcd" package for R (Meyer and Hornik 2006).

Section 2.5 Conclusions

Alternative splicing is clearly an important mode of transcriptome regulation, but analysis by RNA-Seq has challenges at multiple levels, with uncertainty at the alignment stage, at definition of splicing events, and quantification. We have demonstrated the major sources of variation and error prevalent in an RNA-Seq analysis, and shown how using Spanki can mitigate these problems. Implementation of tested filtering steps in Spanki reduces the number of genes called differentially spliced. We argue that without such filtering, variation and error result in over-estimation of splicing differences in RNA-Seq studies.

2.5.1 Implications for RNA-Seq experimental design

Our results provide guidance on the RNA-Seq depth required to analyze splicing and the optimum type of experiments to perform at the design phase of a project. Since splice junctions are a small fraction of the total mapped reads, a logical course of action is to re-sequence at great depth to improve detection. Perhaps counter-intuitively, our results show that greater depth leads to diminishing returns, as the vast majority of sequenced reads originate from highly expressed transcripts.

This has been observed previously in data from human samples (Labaj, Leparc et al. 2011; Tarazona, García-Alcalde et al. 2011). In our experiments, the vast majority of junctions are true positives at 20 million mapped reads, but new junctions detected due to increased coverage at 40 million mapped reads and false positive junctions are equally frequent. We show that with proper curation using Spanki, it is possible to extract real junctions at high sequencing depths (> 100 million real RNA-Seq reads), one can identify rare junctions that are unlikely to be due to experimental error.

The variance we observe between biological replicates suggests that for improving inference on between-sample splicing differences, biological replication in RNA-Seq experiments is essential. Given that junction discovery declines rapidly with great sequencing depth, and that current technology allows us to obtain > 100 million read pairs per lane (Table 2.1, Illumina HiSeq 2000, San Diego, CA), sufficient depth can be obtained for greater than one sample per lane. Multiplexing independent libraries in one lane (Wang, Si et al. 2011) is therefore a good strategy for obtaining adequate depth and biological replication.

2.5.2 Spanki

Sequencing technology is rapidly evolving. One important benefit of Spanki is that each experiment can be modeled, to provide more robust inferences based on the error characteristics of data at hand as sequencing chemistry, devices, and aligners evolve. Spanki integrates error modeling and detection, to provide a more coherent analysis that is adaptable to each experiment. For example, a variety of spliced aligners are available, which offer different strategies for aligning and filtering. Since error profiles may differ by experiment type and sequence characteristics of the reference genome, the Spanki simulation tool allows the comparative assessment of alignment strategies with great specificity. Although the error types we have identified using Spanki are generalizable, a specific assessment of error is a critical component of any comprehensive analysis. For our particular experiments, we have shown that junction detection by TopHat (Trapnell, Pachter et al. 2009) in Drosophila RNA-Seq data is sensitive (> 90% detection at 6x coverage), specific (< 0.5% false positives), and accurate (r = 0.99). Spanki supports results from other aligners that produce output in the standardized BAM format (Li, Handsaker et al. 2009).

Since confidence in an alignment has both quantitative and qualitative characteristics, it is preferable to have the flexibility to use different criteria for different junctions and experiments. For example, if one is studying a splicing mutant altering splicing fidelity, more permissive criteria for aberrant junctions may be required. As another example, if one is interested in recovering minor form introns, additional stringency can be applied. Support Vector Machine (SVM) predictions of valid intron sequence have been used as part of an alignment strategy (QPALMA, (De Bona, Ossowski et al. 2008)), but minor form introns are a much smaller class from which to train. A post-hoc filtering strategy affords greater flexibility to apply more stringent quantitative criteria (such as entropy scores) to putative unannotated minor forms. The filtering criteria generated by Spanki can also be used to assess non-canonical junction detections for which little is known about intron sequence characteristics (for example from HMMSplicer, (Dimon, Sorber et al. 2010)). This added functionality in Spanki fills a crucial gap in curating junction alignments to obtain a high-confidence set of junction calls.

Several analysis tools, such as JUNCbase, (Brooks, Yang et al. 2011) and Splicegrapher, (Rogers, Thomas et al. 2012), produce splicing event definitions into basic categories, but transcript diversity requires a more inclusive classification ontology. Our study shows the utility of a systematic classification system using AStalavista (Sammeth, Foissac et al. 2008). Our classification of events found that 14% of events were of a complex type that did not fall into a basic event category (Black 2003), and which would be excluded from a more narrow definition of events.

In the case of multiple-exon events (such as cassette exons), we found that up- and down-stream connections from the internal cassette exons make it difficult to quantify the event with specificity. This potential for confounding is not accounted for in other count based approaches such as JUNCbase (Brooks, Yang et al. 2011) or DEXseq (Anders, Reyes et al. 2012). Spanki includes a unique feature where junction connections surrounding the event are analyzed, so that events that are potentially affected can be flagged.

When splicing occurs upstream of an alternative promoter, alternate promoter use can be estimated with our approach by using junctions that differentiate these isoforms. However, we find this inference is less reliable than for internal events. In simulations, we found PSI estimates for alternative first exon events as a class to have more variance than internal splicing events, especially in cases where the first exon is short, which gives less territory for read sampling. Experiments that specifically target the 5 prime end of transcripts, such as cap analysis of gene expression (CAGE) are better suited for detailed analysis of alternative promoter use (Takahashi, Kato et al. 2012).

2.5.3 Sex differential splicing in Drosophila heads

Alternative splicing is widespread in mammals, with estimates suggesting that transcripts of 95% of all genes are alternatively spliced in humans (Pan, Shai et al. 2008; Wang, Sandberg et al. 2008), but is less prominent in Drosophila (%40 of annotated genes) (Graveley, Brooks et al. 2011). In head tissue, we detect alternative splicing (where a gene produces > 1 splice variant in either sex) in 16% of all genes. However, we also detect extensive low-abundance junctions not incorporated into transcript models in an additional 13% of genes. In mammals, it has been proposed that a large class of processed transcripts represent low abundance 'noisy' splicing (Pickrell, Pai et al. 2010). We detect low abundance splices in transcripts of 35% of genes, suggesting the presence of a similar phenomenon in flies. This poses a difficult problem for annotation of gene models as it is difficult to determine if a high confidence junction was obtained from an important, but rare, splicing event; or from a tolerated biological error.

Several high-throughput studies have attempted to globally quantify sex-differential splicing. One microarray study examined head tissue specifically and identified 12 genes encoding transcripts that were sex-differentially spliced (McIntyre, Bono et al. 2006). However, limitations of microarray platforms without complete probesets prevented a full interrogation of all genes in that study. RNA-Seq experiments have no such limitation, and a recent estimate of sex-differential splicing in head tissue from RNA-Seq experiments suggested that there are 1,370 sex-differentially expressed transcripts (Chang, Dunham et al. 2011). Our results suggest that this figure is at least 10- to 100-fold overestimated. Strikingly, the genes showing sex-biased steady-state transcription profiles are not the same genes that show sex-biased splicing. Like sex-biased splicing, we find that sex-biased gene expression is modest in heads (19 genes).

Many of the sex-biased splicing events we observed could be important for sexual behavior. *Reticulon-like 1* (*Rtnl1)* had significant differences at several pairwise defined alternative first exons. *Rtnl1* is a membrane protein localized to the endoplasmic reticulum (ER) (Wakefield and Tear 2006) and has a role inter-male aggressive behavior (Edwards, Zwarts et al. 2009), olfactory response (Sambandan, Yamamoto et al. 2006), and motor axon development (O'Sullivan, Jahn et al. 2012). Another gene with sex-differential skipped exons, multiplexin, is involved in motor axon guidance (Meyer and Moussian 2009), although without a known link to behavior. We detected sex-differential regulation in transcripts encoded by the found in neurons (*fne*) gene, which encodes a member of the embryonic-lethal abnormal vision (ELAV) gene family of RNA-binding proteins (Samson and Chalvet 2003; Pascale, Amadio et al. 2008). Wildtype *fne* is required for robust male courtship behavior (Zanini, Jallon et al. 2012).

<u>*Section 2.6 Materials and methods*</u>

2.6.1 Generation of RNA-Seq data

Sample descriptions and detailed methods for generating RNA-Seq data are provided in Gene Expression Ominibus (GEO) accessions (GSM928376, GSM928377, GSM928383, GSM928384, GSM928392, and GSM928393). Briefly, RNA samples were prepared from adult Drosophila heads of each sex, for wild type flies. The wild-type (WT) strain was a white^{1118}, Canton-S (B) isogenic stock obtained from Trudy Mackay (North Carolina State University, Raleigh, NC) (Edwards, Rollmann et al. 2006; Yamamoto, Zwarts et al. 2008). Flies were grown at low density and were aged for 7 days post-eclosion and flash frozen on dry ice. Heads were dissected from frozen flies with forceps on chilled ceramic plates. Two independent collections were performed for each sex, and cDNA libraries were prepared. Fly heads were transferred on dry ice into pre-cooled screw-capped 1.5 ml. tubes with 1.4 mm diameter ceramic beads, and total RNA was extracted using Trizol (Life Technologies (Invitrogen), Grand Island, NY, USA) Mechanical homogenization was for two periods of for 30 seconds at 6500 rpm in a Precellys24 homogenizer (Bertin Technologies, Aix-en-Provence, France), cooled to between 9-12^{o} C with liquid nitrogen. Exogenous controls from the External RNA Control Consortium (ERCC, pool 15) were spiked in at 1% concentration prior to library construction (Jiang, Schlesinger et al. 2011). Paired-end sequencing was performed on GAII or HiSeq instruments from Illumina (San Diego, CA) for 76 cycles for each read mate.

2.6.2 Read alignments

Reads that passed Chastity base-calling filtering (Illumina CASAVA pipeline 1.6.47.1) were used for further analysis. The default Chastity score threshold used was > 0.6, and is defined as the ratio of the highest of the four (base type) intensities to the sum of highest two (Illumina, San Diego, CA). Mapping was performed using TopHat v1.4.1 (Trapnell, Williams et al. 2010), with Bowtie v0.12.7 (Langmead, Trapnell et al. 2009), and samtools 0.1.12a (Li, Handsaker et al. 2009), and parameters “-g 1 –solexa1.3-quals, -i 42.” A reference annotation (Ensembl release 67, corresponding to Flybase 5.39) was also supplied in GTF format with the -G option. Briefly, TopHat aligns reads to a reference transcriptome first, setting aside unaligned reads as putative splice-junction spanning reads. It then builds a new reference composed of joined exon sequence of all possible putative joins between islands of genomic coverage identified in the first-pass alignment. Supplying the reference annotation guaranteed inclusion of annotated junctions into the dynamically-generated exon-join reference, but does not limit the algorithm to annotated junctions. TopHat performs these alignments with the Burrows-Wheeler aligner Bowtie (v.0.12.7) (Langmead, Trapnell et al. 2009).

We used *D. melanogaster* genome release 5 (Celniker and Rubin 2003; BDGP 2006), as obtained from the UCSC genome browser (excluding "chrUextra”)(Kent, Sugnet et al. 2002), for mapping. We also appended sequence for 96 exogenous controls to the genomic reference (Jiang, Schlesinger et al. 2011). We used the reference annotation obtained from Ensembl (release 67, May 2012; imported from Flybase release 5.39, July, 2011 (McQuilton, St Pierre et al. 2012)). A minor modification was made to remove only the antisense transcripts of *modifier of mdg4* (FBgn0002781), since the presence within a gene of transcripts on both strands caused fatal errors in downstream analysis tools.

2.6.3 Feature detection and quantification

To produce estimates of gene and transcript level abundance, we quantified based on both full-length transcript assemblies and on discrete counts within annotated genomic boundaries, as each approach has different strengths (Anders and

Huber 2010; Trapnell, Williams et al. 2010). Cufflinks (Trapnell, Williams et al. 2010) (v.2.0.2) was used for generating abundance estimates of full-length isoforms. Briefly, Cufflinks performs assembly of putative transcripts de novo, or against a supplied annotation. It then generates maximum likelihood estimates of the expected number of fragments in the sample originating from your gene of interest in units of expected fragments per kilobase of transcript per million mapped reads (FPKM). We produced FPKM values with Cufflinks using the Ensembl annotation described above. To maximize sensitivity, we turned off minimum isoform fraction filtering, and set the minimum intron size to 42 (the size of the smallest annotated intron).

Transcriptional differences between samples were analyzed using transcript abundance estimates from Cuffdiff (Trapnell, Williams et al. 2010), and counts of read alignments within genes in HTseq/DEseq (Anders and Huber, 2010). Briefly, Cuffdiff takes as input transcript abundance estimates from Cufflinks. Gene level abundance comparisons are made by t-test, where the variance term is estimated from the beta negative binomial distribution. Cuffdiff v.2.0.2 was used to compare between samples, using upper quartile normalization (to improve robustness of calls in low-expressed transcripts), and setting “max-bundle-frags” very high (50E06), to ensure that very highly expressed features were not excluded. To provide alternative quantifications and comparisons, we used HTseq (Anders and Huber 2010) to generate simple counts of reads that fall within discrete features. The “htseq-count” program in HTseq v.0.5.3, with the conservative “union” mode and default parameters, was used to generate counts. We used the R package DESeq (v.1.8.3) to test for differential expression by modeling variance with a negative binomial model, while adjusting variance estimates by expression intensity (Anders and Huber 2010). "Variance outliers" were identified as contrasts where the maximum residual variance is > 15. This value is exceeded in ~2% of all genes. These outliers are removed from our final list of differential gene expression.

2.6.4 Quantitative Real Time PCR

Total RNA was extracted with TRIzol reagent (Life Technologies (Invitrogen), Grand Island, NY, USA) from heads of 7 day old *white*^{1118}, *Canton-S (B)* isogenic females and males. One µg of total RNA was subjected to DNase treatment (Promega, Madison, WI, USA) followed by reverse transcription, using the random primer of the Transcriptor First Strand cDNA synthesis kit according to the manufacturer's instructions (Roche Applied Science, Indianapolis, IN, USA). Quantitative real-time PCR was run for 2 independent cDNA preparations, each with duplicate quantification (4 measures per primer pair). cDNA corresponding to 12.5 ng of total RNA was amplified with Fast SYBR Green Master Mix (Applied Biosystems, Carlsbad, CA, USA; 10µl reaction) in a StepOne Real-Time PCR machine (Applied Biosystems, Carlsbad, CA, USA). The qPCR program: initial activation was performed at 95°C for 20 seconds followed by 40 cycles. DNA strands were separated at 95°C for 3 seconds followed by an annealing at 60°C for 30 seconds. Then the melting curve was generated ranging from 60°C to 95°C with an increment of 0.5°C each 5 seconds. *Act5c* (*Actin 5C*) was used as a control. Primers were designed with the web interface of the NCBI Primer-Blast software (Rozen and Skaletsky 2000). All qPCR primer products were verified as possessing a single peak during amplification, confirming that only a single product was being produced. All amplification products were analyzed by agarose gel electrophoresis and produced fragments of predicted sizes (not shown). The relative transcript level was calculated using the cycle threshold value (Ct) by the method of $2^{-\Delta Ct}$, where $\Delta Ct = Ct_{transcript} - Ct_{Act5c}$. qPCR data are provided for each primer pair, after normalization to junction coverage of the mutually exclusive isoform.

Section 2.7 Author contributions

A version of this chapter was submitted for publication with the following authors:

Sturgill D, Malone JH, Sun X, Smith HE, Rabinow L, Samson ML, Oliver B (2012). A Junction-based Splicing Analysis Kit (Spanki) for RNA-Seq data. (Submitted manuscript)

I developed and implemented all computational methods and was the primary author of the text. Leonard Rabinow, Marie-Laure Samson, and Brian Oliver also substantially co-wrote and edited the manuscript. The project as a whole was conceived by myself, Leonard Rabinow, Marie-Laure Samson, and Brian Oliver.

Leonard Rabinow and Marie-Laure Samson provided the biological samples, and John H. Malone performed Illumina library construction and performed RNA-Seq experiments along with Harold Smith. Xia Sun designed primers and performed qPCR experiments.

Chapter 3: Characterization of post-transcriptional regulator mutants

Section 3.1 Introduction

The somatic sex determination cascade in Drosophila is regulated by differential splicing, from an autoregulatory loop for the master switch at initiation (*Sxl*) to the production of different isoforms of DNA-binding transcription factors at the termination (*dsx* and *fru*) (Venables, Tazi et al. 2011). These splicing differences are tightly regulated and highly sex-specific, and dependent on multiple components of the splicing machinery. Downstream targets of sex-differential splicing are not widely known, and may consist of many subtle differences regulated in specific cellular and developmental contexts, along with multilayered and interdependent connections to other regulators. One way we can gain further insight into regulated dimorphic splicing is by finding transcripts that differ between the sexes in their response to mutation of post-transcriptional regulators.

There are many examples of post-transcriptional regulators that play a role in sexual differentiation. One example is *Darkener of apricot* (*Doa*), a protein kinase of the LAMMER/Clk family that phosphorylates serine - arginine rich proteins (SR proteins) such as transformer (TRA) transformer-2 (TRA2) and RBP1 (Rabinow and Samson 2010), which are critical for the regulation of sex-differential splicing. Mutants of *Doa* exhibit hypophosphorylation of these proteins and exhibit sex-reversal phenotypes (Rabinow and Samson 2010). *Doa* also produces a sex-specific isoform from an alternative promoter proximal to a DSX binding site, suggesting that

Doa is part of the ancient sex determination network along with the conserved *tra* – *dsx* axis (Kpebe and Rabinow 2008; Rabinow and Samson 2010). *Doa* is broadly expressed (detected in all tissues examined), and *Doa* mutants are recessive lethal, including in individual cell clones (L. Rabinow, personal communication).

Another post-transcriptional regulator is *found in neurons* (*fne*). *fne* is a member of the embryonic-lethal abnormal vision (ELAV) gene family of RNA-binding proteins, involved in the regulation of RNA metabolism (Samson and Chalvet 2003; Pascale, Amadio et al. 2008). The ELAV family is well conserved across many eukaryotic lineages, and is required for neuronal differentiation and maintenance. RNA-Seq experiments show that transcripts of *fne* exhibit sex-differential regulation at an alternative donor site (Chapter 2), and wildtype *fne* is required for robust male courtship behavior (Zanini, Jallon et al. 2012). Mutants of *fne* also show brain anatomy defects (fusion of mushroom body neurons), and preliminary experiments show fne^{null} males exhibit little general aggressive behavior (M. Samson, personal communication). Expression of *fne* occurs only in neurons, and mutants are viable and fertile (Zanini, Jallon et al. 2012).

To better understand the function of these genes as well as place them in context of these genes in sexual differentiation, we performed RNA-Seq analysis on mutant lines, and analyzed the consequences of disrupting these loci on the transcriptome of *Drosophila* heads. We performed a thorough characterization of the regulatory effects of these mutants, including changes to gene expression, relative isoform abundances, intron retention, and aberrant splicing.

Section 3.2 Results: Gene expression

We generated two independent replicate pools of RNA for library construction, for female and for male adult heads, for two strong hypomorphic heteroallelic mutants of *Doa* (*Doa*^{EMS} and *Doa*^{HD}) (Rabinow and Samson 2010), and one null mutant allele of *fne*, and (See Materials and methods).

Although neither *Doa* nor *fne* regulate transcription directly, downstream effects of the transcripts they regulate post-transcriptionally may be evident in the global response of the transcriptome. To examine the transcriptional response of each mutant, we compared gene expression for each mutant vs wild type, by generating counts of reads in each gene, and testing for differences with DESeq (Anders and Huber 2010) (Figure 3.1).

We observed greater transcriptional differences in *Doa* mutants relative to wild type than in *fne* mutants, suggesting that pleiotropic effects of targets with disrupted splicing in these lines are extensive (Figure 3.1). In each *Doa* mutant vs wt comparison, a greater number of targets were downregulated in mutants, with the exception of *Doa*^{HD} males, which showed much greater upregulation in mutants (126 genes upregulated vs 27 genes downregulated).

Since *Doa* mutants affect sexual differentiation, we examined the transcriptional effects on genes with sex differential expression in wildtype. Rather than reversals of sex bias, we observed complex responses to *Doa* mutation by sex. For example, we found significant downregulation of the Turandot family of genes in *Doa* males. The Turandot family,a set of eight genes involved in the humoral stress response, exhibit male biased expression in wildtype males. In contrast, several odorant binding proteins also exhibit male-biased expression in wildtype, but they have different responses to *Doa* mutation. One of these genes (*Obp19b)* is downregulated in *Doa* males and unchanged in *Doa* females, while *Obp99b* is upregulated in both *Doa* males and females. Genes with significant female biased expression in wildtype, including *yp1*, *yp2*, *yp3*, and *fit*, were unchanged in *Doa* females, but downregulated in *Doa* males. These results suggest that complex network interactions may be at play in regulating these targets of sex-differential transcription.

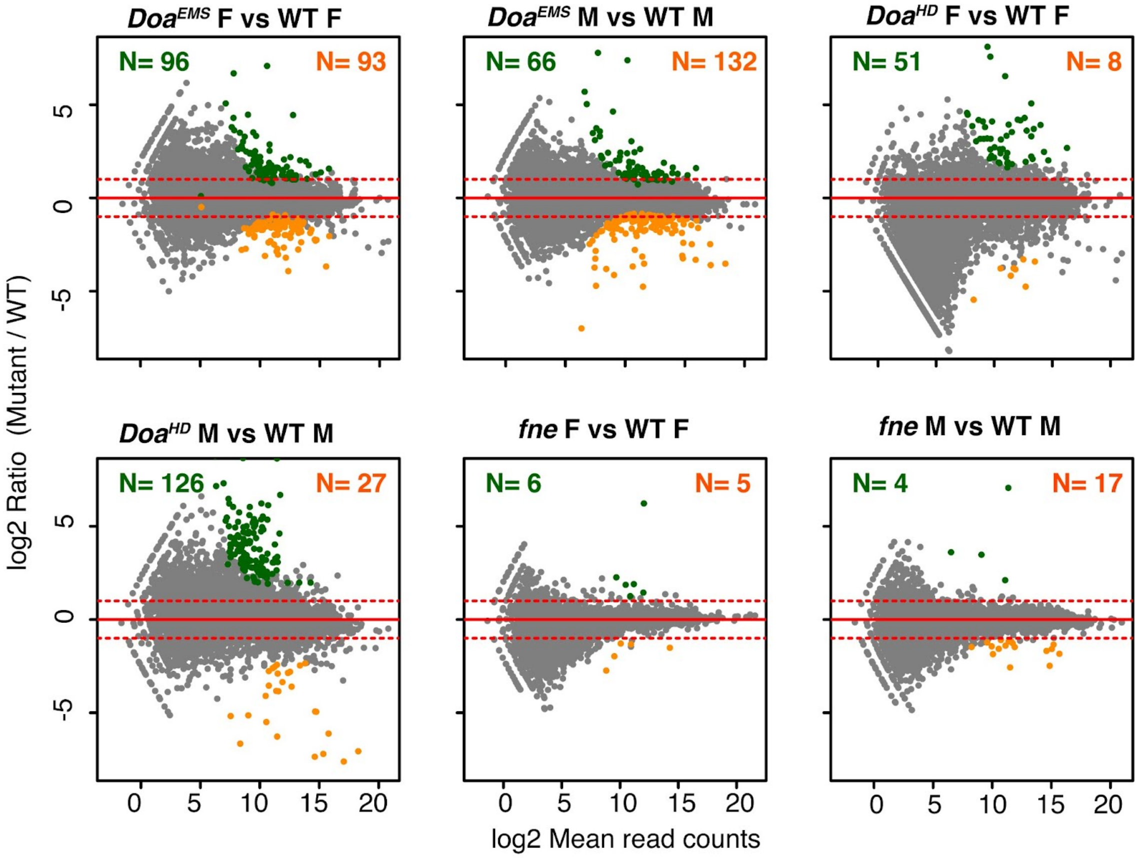


Figure 3.1: Gene expression differences in mutant vs wild type (WT) in each sex. Each panel is an "MA plot," showing the ratio of gene expression differences (log2 ratio of mutant/wt expression) on the y axis, and the average total abundance (log2 mean read counts) on the x-axis. Inset are numbers of genes differentially expressed in each comparison, at 5% false discovery rate (FDR). Female samples are denoted "F" and male samples are denoted "M" in this figure and throughout this chapter.

Section 3.3 Results: Splicing

3.3.1 Effects on splicing events

RNA-Seq experiments capture detailed information about processed transcripts in an RNA sample, allowing us to examine the effects on splicing in *Doa* and *fne* mutants. We used the Splicing Analysis Kit (Spanki, see Chapter 2) to characterize the post-transcriptional effects of these mutants at multiple levels, by comparing pairwise events and individual introns.

We first generated proportion spliced-in values (PSI) for all annotated splicing events and clustered them, to look for general patterns of the metric in each sample. This metric expresses the relative proportion of two mutually exclusive splice forms (an "inclusion" and "exclusion" form). We found that there was not extensive reversal of direction of inclusion / exclusion preference. We also found that different *Doa* mutant alleles of the same sex clustered together, but apart from wildtype (Figure 3.2). Each wildtype sex clustered with the *fne* sample of the same sex, while the sample from each *Doa* allele clustered together by sex. These results show that PSI values change broadly in response to *Doa* mutation, but this response is different between the sexes.

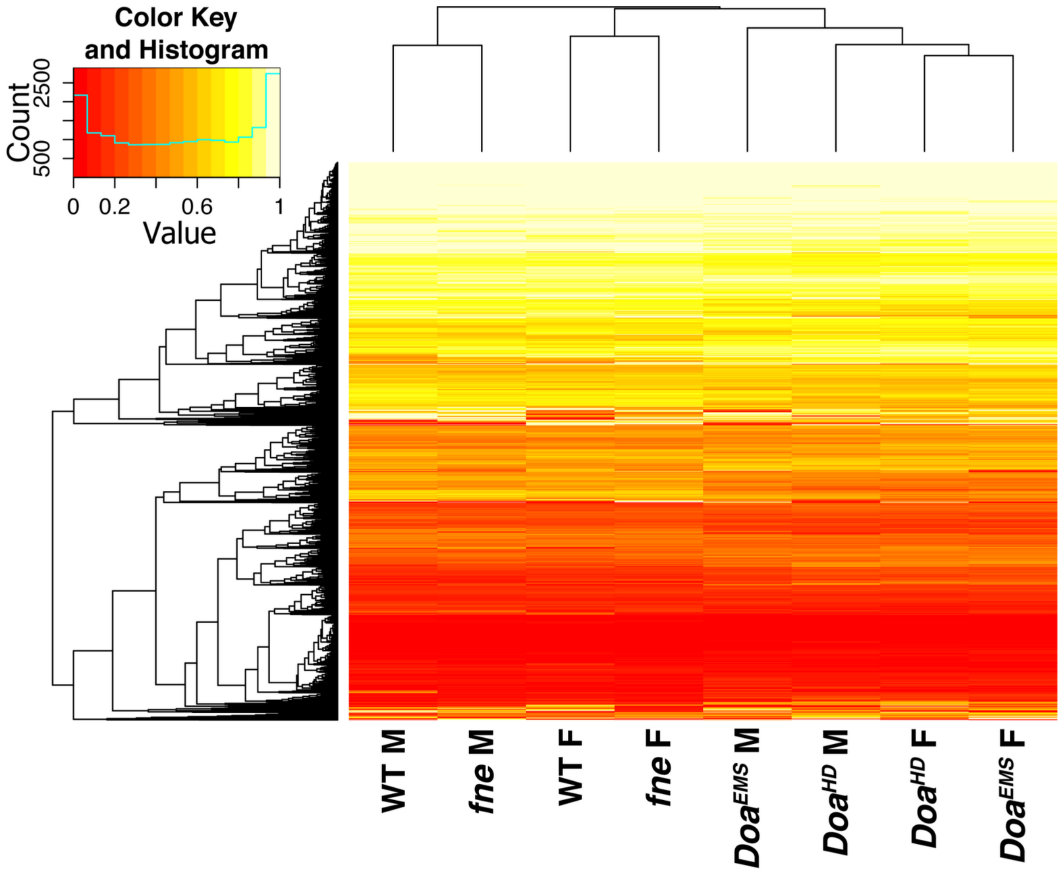


Figure 3.2: Clustered heatmap of proportion spliced in (PSI) for splicing events in each sample. Rows and columns are both hierarchically clustered.

Differences in splicing regulation in response to mutation may involve subtle differences in either sex or both sexes, and therefore may be difficult to resolve. Because of this, multi-way comparisons may be useful. Figure 3.3 illustrates the example of the effect on splicing in the gene *jetlag*. This gene is involved in circadian rhythm and behavior, and exhibits subtle sex-differential splicing at an alternative acceptor event. Both sexes favor the exclusion form, but this bias toward the exclusion form is greater in females. In *Doa* mutants of both sexes, the bias toward exclusion is diminished. The net effect in a *Doa* female vs *Doa* male comparison is to have an equal bias toward exclusion in both sexes, while the subtle sex difference in these proportions are still visible in the *fne* female vs *fne* male comparison (Figure 3.3). A splicing effect in *jet* is consistent with the phenotypes of *Doa* mutants, which are arrhythmic (Francois Rouyer, personal communication).

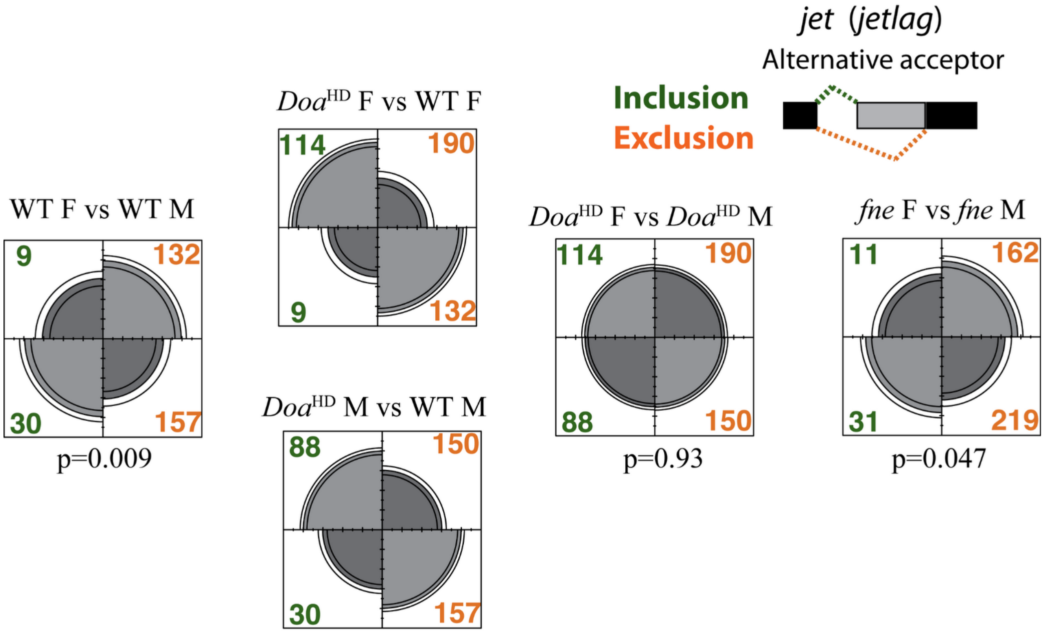


Figure 3.3: Fourfold plots of an alternative acceptor event in the gene *jetlag*, in wild-type and mutant. The counts for each inclusion or exclusion form are inset, and the significance values from a Fisher's exact test are given below the wildtype female vs wildtype male, and mutant female vs mutant male comparisons.

We compared the response to *Doa* mutation in each sex, to infer the nature of the regulated event (Figure 3.4). In results for *fne* and *Rtnl1*, the sex difference in splicing was removed *Doa* males, but not in *Doa* females. This suggests that these events are regulated in wild-type females, and follow a default splicing pattern in wild type males. This is similar to components of the sex determination pathway, which produce a default splice in males in the absence of TRA (Figure 3.4). However, the *Rtnl1* splicing event is an alternative first exon, so this difference is more likely regulated transcriptionally than by splicing. It is therefore unlikely to be affected directly by *Doa*, suggesting that a trans-acting transcriptional regulator regulating *Rtnl1* alternative promoter usage is affected by *Doa*.

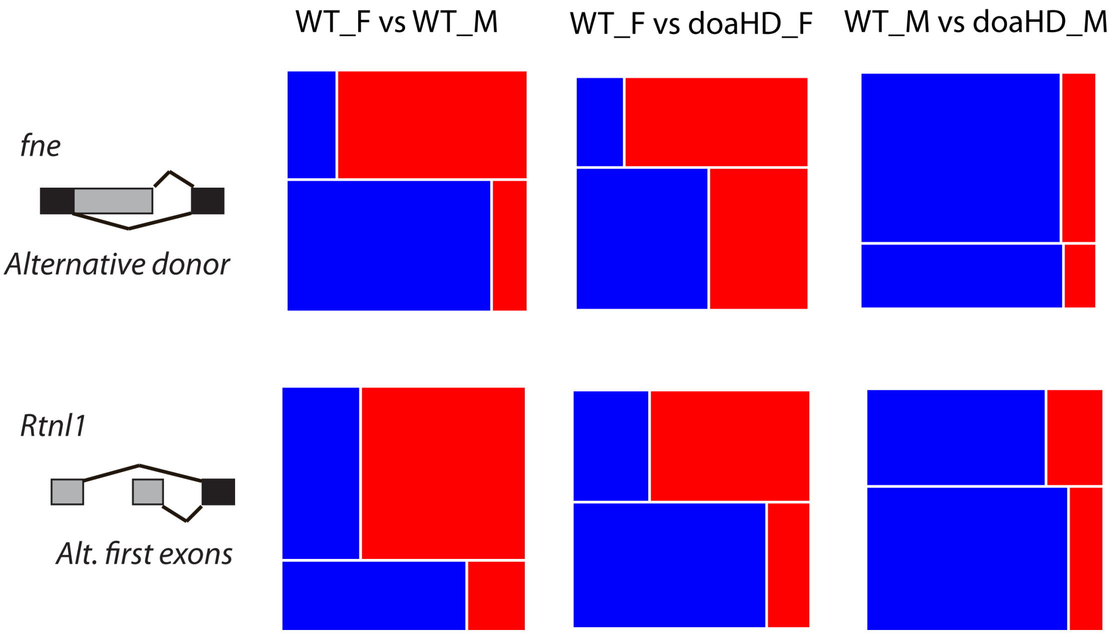
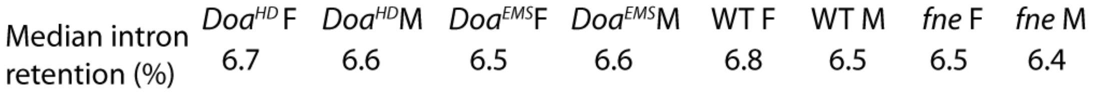


Figure 3.4: Example of changes in sex-differential splicing in mutants. Changes observed in *Doa* mutants to splicing events in transcripts of *fne* and *Rtnl1*. For each event, results in females are given in the first row, and males in the second row, using a mosaic plot. Blue and red colors represent the splice form predominate in wild-type males and females, respectively.

3.3.2 Effects on intron retention

The Spanki toolkit generates junction level results that reflect splicing in RNA-Seq data that is not captured from an analysis of annotated splicing events. For example, Spanki calculates intron retention for every splice junction, not just for those with an annotated retained intron isoform. We used these results to estimate ‘background’ intron retention rates, detect transcriptome-wide intron retention differences between samples, and identify subtle changes in intron retention between samples.

We first examined the possibility of a global effect on intron retention in all mutant samples. To estimate global intron retention, we selected junctions that are in constitutively spliced genes, have no joins to novel alternative donors or acceptors, and have coverage > 10, for all samples in the highest coverage run (Run 63). We found no significant difference across samples, and the average median intron retention was 13.1% (Figure 3.5). These results suggest there is no significant global intron retention difference in mutant samples.



	Doa^{HD} F	*Doa*^{HD}M	*Doa*^{EMS}F	*Doa*^{EMS}M	WT F	WT M	*fne* F	*fne* M
Median intron retention (%)	6.7	6.6	6.5	6.6	6.8	6.5	6.5	6.4

Figure 3.5: General intron retention. Median intron retention values for all constitutive splice junctions in each sample

Individual introns may be differentially retained by sex and serve a regulatory role. The retained intron may disrupt the reading frame of the processed transcript or introduce a premature stop codon, or influence regulation by another means such as affecting transcript stability. We examined retention differences in individual introns of specific genes, to see if they had a difference with regard to sex or mutant allele.

We also analyzed intron retention estimates from Spanki at each intron in *E2f2*, a DNA binding transcription factor that has mutltiple roles in development (Dimova, Stevaux et al. 2003). This analysis revealed a significant difference in retention for the 2nd intron in wild-type males, that is not observed in *Doa* mutants (Figure 3.6). Retention of this intron would disrupt the reading frame of the *E2f2* transcript. Additionally, the fourth intron has many differences between samples, but has much more variance (Figure 3.6). Variance in read coverage at the most three prime end has been observed as a general effect previously in RNA-Seq experiments (Wang, Gerstein et al. 2009). Nevertheless, a highly significant increase in intron retention is observed at the intron in *Doa*^{EMS} males. Although retention of this intron does not alter the reading frame or include an in-frame stop codon, it may still have a regulatory role, as splicing of *E2f2* transcripts has been shown to be altered in *Doa* mutants (Rasheva, Knight et al. 2006; Ambrus, Rasheva et al. 2009).

An RNAi screen revealed E2F2 target genes that are sex-differentially transcribed and involved in reproductive phenotypes including courtship behavior (Dimova, Stevaux et al. 2003). These results suggest that *Doa* is involved in indirect regulation of these downstream targets by modulating intron retention in transcripts of *E2f2*.

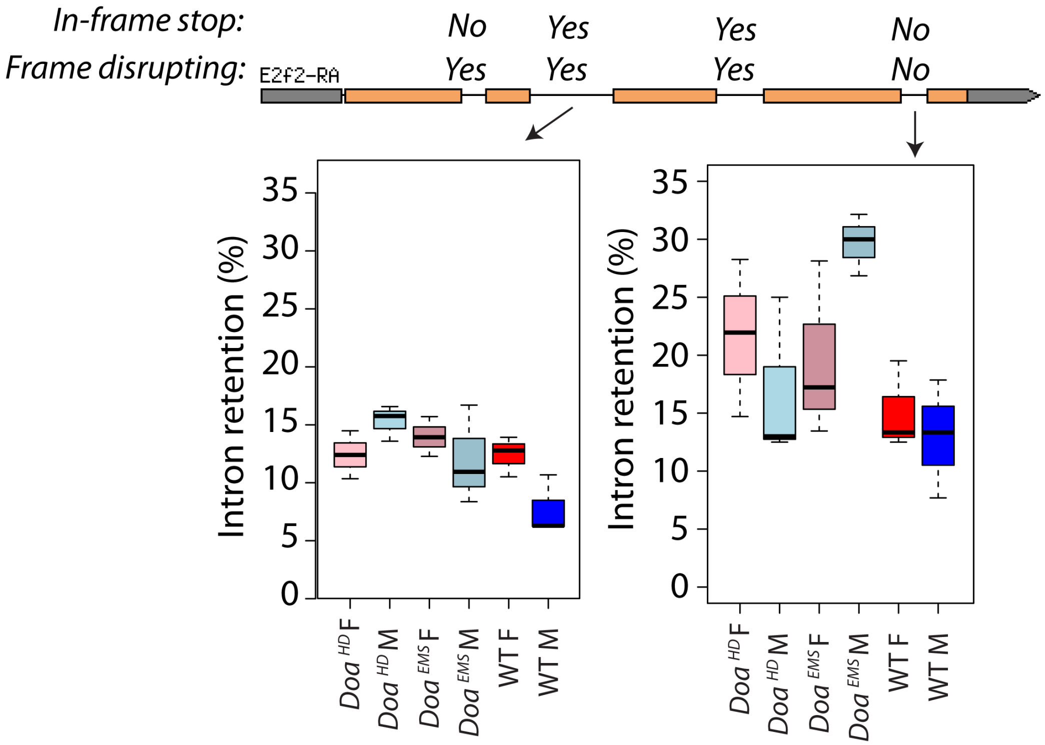


Figure 3.6: Detailed analysis of *E2f2* introns. The most three-prime intron has the greatest variability. Each box represents three values (for the three replicates). The middle line is the middle value, the filled box is the range from average of the two lowest to average of the two highest. The dotted lines extend to the full range of the three values.

We also observed differences in intron retention in transcripts of *sex specific enzyme 2* (*sxe2)* - a gene that encodes a phospholipase, and that is significantly upregulated in wild-type males (Fujii and Amrein 2002) (see also Chapter 2). There are two introns in this gene that are both frame preserving, but each has an in-frame stop codon. Both introns are retained slightly less in wild-type male, while the first intron is retained much more frequently in *Doa* females (Figure 3.7). These results suggest a role for *Doa* in negative regulation of this transcript in females which is already downregulated transcriptionally.

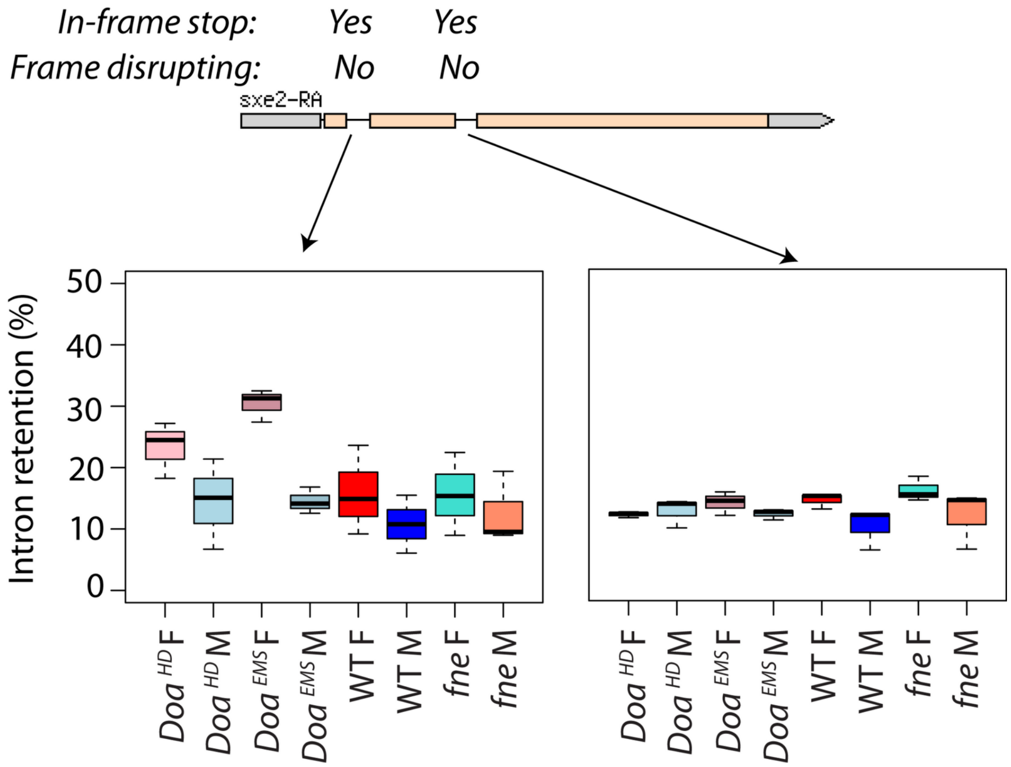


Figure 3.7: Intron retention in *sxe2* transcripts. Boxplots of all replicates (see Figure 3.6 for additional explanation)

3.3.3 Aberrant splicing

Unannotated junction detection may be the result of novel discovery of splices found in wildtype, or an indication of an aberrant join in mutants. We detected an increase of 13.5% to 63% more unannotated junctions in *Doa* mutants than in wildtype. The largest increase (63%) was in the *Doa*^{HD} female vs wildtype female comparison. There was no overall increase in unannotated junction detection in *fne* mutants. To discern aberrant splicing from alignment artifact is difficult, and the criteria often used filter false positives is based on known biology in wildtype (Chapter 2). However, the fact that these junctions were not detected in matched wildtype samples, where the parent RNA population is not expected to be radically different, suggests that alignment artifact is an unlikely cause. We further examined the characteristics of these junctions that could indicate aligner error, and found that removing repeat-induced aligner errors, and alignments to minor-form introns removed only a small portion of these detections (for a 52% increase in *Doa*^{HD} females vs wild-type males).

We examined the detection of aberrant junctions in *dsx* transcripts, to see how this critical transcriptional regulator was affected (Figure 3.8). We observed four aberrant splices with this locus. Two of these were joins of annotated exons that are not seen in wildtype (exon 1 to exon 4, and exon 4 to exon 6). Another join was detected between intron space and intergenic space on the opposite strand proximal to exon 1, which was detected only in *fne* mutants. We can not exclude that these reads arose from a neighboring transcript (the three-prime end of *CD98hc*, and also transposable element, are ~7kb from this genomic location). Finally, we observed a join with a ~300 bp gap with exon 3, upstream of the start codon. These alignments were detected in more than one *Doa* replicate and with multiple reads. These junctions possessed GT/AG motifs, but did not contain sequence characteristics of a true donor site or a polypyrimidine tract (See Chapter 2, figure 2.3). These results are informative of the mutational lability of splicing regulation, and suggest that alterations of splicing factors can lead to the generation of new processed transcripts, not just changes in amounts.

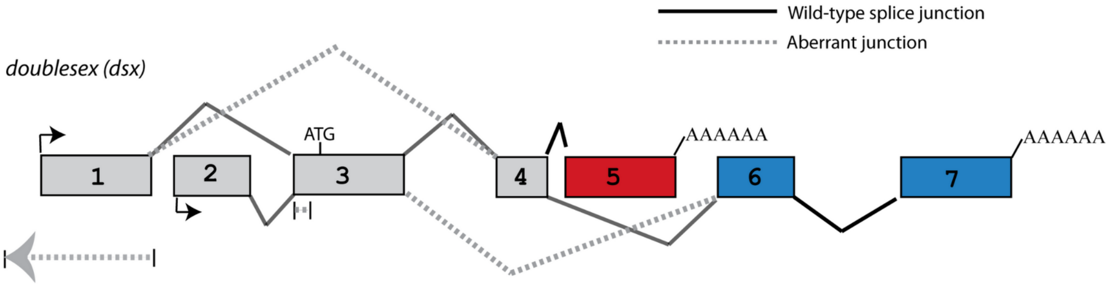


Figure 3.8: Aberrant splice junctions in *dsx* transcripts detected in mutants. A cartoon of the *dsx* gene model is shown, with the female specific exon shown in red, and the male specific exons in blue. The aberrant junctions are detected only in mutants, and are shown in grey dotted lines.

Section 3.4 Discussion and conclusions

Post-transcriptional regulation is essential to normal development and physiology. We observe extensive effects consequent to disrupting splicing by *Doa* mutation including pleiotropic gene expression changes, alterations in proportions of alternative splicing variants, changes in retention of specific introns, and aberrant splicing. We observe all these phenomena in *fne* mutants as well, but much more modestly. Our results display the huge potential of RNA-Seq analysis to provide a complete and multi-faceted picture of differences between transcriptomes, and show how the Spanki toolkit can facilitate this analysis.

We have demonstrated the gain and loss of sex differences in splicing in *Doa* mutants as well. These results suggest that *Doa* is operating within sex determination pathways. This is consistent with the hypothesis that *Doa* is a part of an ancient regulatory pathway that predates the acquisition of *Sxl* as the master sex-determining switch in *Drosophila* (Rabinow, 2010). Additional experiments are ongoing to provide validation of these effects and to demonstrate phenotypes related to sexual development and behavior. These results will allow us to illuminate regulatory connections to other genes, and place *Doa* in the context of a broader and interacting sex determination network.

Section 3.5 Materials and methods

Sample descriptions and detailed methods for generating RNA-Seq data are provided in Gene Expression Ominibus (GEO) accession GSE37811. Briefly, RNA samples were prepared from adult *Drosophila* heads of each sex, for wild type and and heteroallelic mutants for two alleles of *Doa* (*Doa*$^{\mathrm{HD}}$ /*Doa*^{DEM} and *Doa*$^{\mathrm{EMS2}}$/*Doa*^{DEM}), and for *fne*. The wild-type (WT) strain was a white^{1118}, Canton-S (B) isogenic stock obtained from Trudy Mackay (North Carolina State University, Raleigh, NC) (Edwards, Rollmann et al. 2006; Yamamoto, Zwarts et al. 2008). Animals were cultured in vials at low density, mixed-sex, 25° C and under 24 hour light and in order to keep their behavior and transcript populations as close to normal as possible and to ablate circadian rhythm. Flies were aged for 7 days post-eclosion and flash frozen on dry ice. Heads were dissected from frozen flies with forceps on chilled ceramic plates, with 500 heads per sample. Two pools of RNA were prepared for each sample (one for *Doa*$^{\mathrm{HD}}$ /*Doa*^{DEM} females, due to low viability, to provide independent biological replicates, and cDNA libraries were prepared. Fly heads were transferred on dry ice into pre-cooled screw-capped 1.5 ml. tubes with 1.4 mm diameter ceramic beads, and total RNA was extracted using Trizol (Life Technologies

(Invitrogen), Grand Island, NY, USA). Mechanical homogenization was for two periods of 30 seconds at 6500 rpm in a Precellys24 homogenizer (Bertin Technologies, Aix-en-Provence, France), cooled to between 9-12^{o} C with liquid nitrogen. Exogenous controls from the External RNA Control Consortium (ERCC, pool 15) were spiked in at 1% concentration prior to library construction (Jiang, Schlesinger et al. 2011). Paired-end sequencing was performed on GAII or HiSeq instruments from Illumina (San Diego, CA) for 76 cycles for each read mate.

Data processing including alignments were performed as described in Chapter 2. Gene level abundances and splice junction coverage were also quantified as described in Chapter 2.6. Hierarchical clustering was performed using the 'hclust' function in R, called by the 'heatmap.2' function in the gplots package (Gregory R. Warnes, http://cran.r-project.org/web/packages/gplots/).

Section 3.6 Author contributions

I developed and implemented all computational methods and was the primary author of the text in this chapter. The project as a whole was conceived by myself, Leonard Rabinow, Marie-Laure Samson, and Brian Oliver.

Leonard Rabinow and Marie-Laure Samson provided the biological samples, and John H. Malone performed Illumina library construction and performed RNA-Seq experiments along with Harold Smith.

Chapter 4: Comparisons within the *Drosophila* genus

Section 4.1 Introduction

Phylogenetic divergence is driven by genetic changes, and a major goal of evolutionary genetics is to understand the mechanisms by which these changes create new phenotypes. Commonly, this discussion revolves around mutation that changes protein coding sequence. However, gene regulation also diverges, and is hypothesized to play a major role in species divergence (Romero, Ruvinsky et al. 2012).

Evolutionary divergence of gene deployment can arise from many mechanisms, including changes to cis-regulatory elements, alterations to trans-factors, and sequence divergence of splicing enhancers. Several comparative expression analyses have been performed and lineage specific divergence has been identified, but these studies have been primarily on the level of transcription, rather than splicing (Romero, Ruvinsky et al. 2012).

Sexual characters are often under selective pressure at the level of coding sequence and expression, which makes sexual dimorphism an excellent model to study how transcriptomes diverge (Ellegren and Parsch 2007). Within eukaryotes, there are a diverse variety of sex-determination regulatory networks (Williams and Carroll 2009), suggesting that these pathways are evolutionary labile. Gene regulation is also evolutionarily labile, and previous studies have shown that sex-differential transcription diverges over time (Ellegren and Parsch 2007; Zhang, Sturgill et al. 2007; Romero, Ruvinsky et al. 2012).

To examine the divergence patterns of transcription and splicing, we generated RNA-Seq data from multiple species of the Drosophila lineage (Figure 4.1), and examined differences in their transcriptomes, at the level and transcription and splicing.

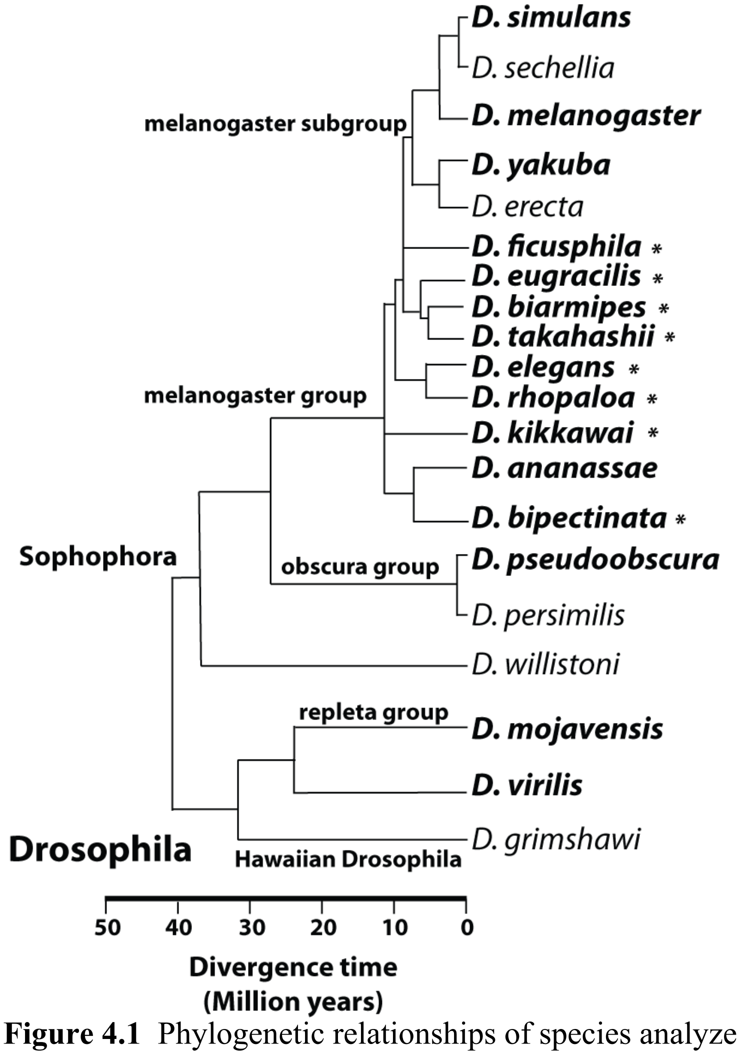


Figure 4.1 Phylogenetic relationships of species analyzed for expression differences. RNA-Seq data from the species in bold are analyzed. Newly sequenced species that are not part of the original 12 sequenced species (Clark, Eisen et al. 2007) are asterixed. Adapted from Flybase, with new species added according to published estimates of phylogenetic relationships (Kopp 2006; Piano and Cherbas 2010).

Section 4.2 Results

4.2.1 Gene expression differences

We generated pools of RNA from sexed whole adults of seven species of Drosophila, broadly sampling the phylogeny of the sequenced species. We produced gene level expression estimates using Cufflinks (Trapnell, Williams et al. 2010) in units of fragments per kilobase per million mapped reads (FPKM), and tested for sex-differential expression in each species (Figure 4.2)

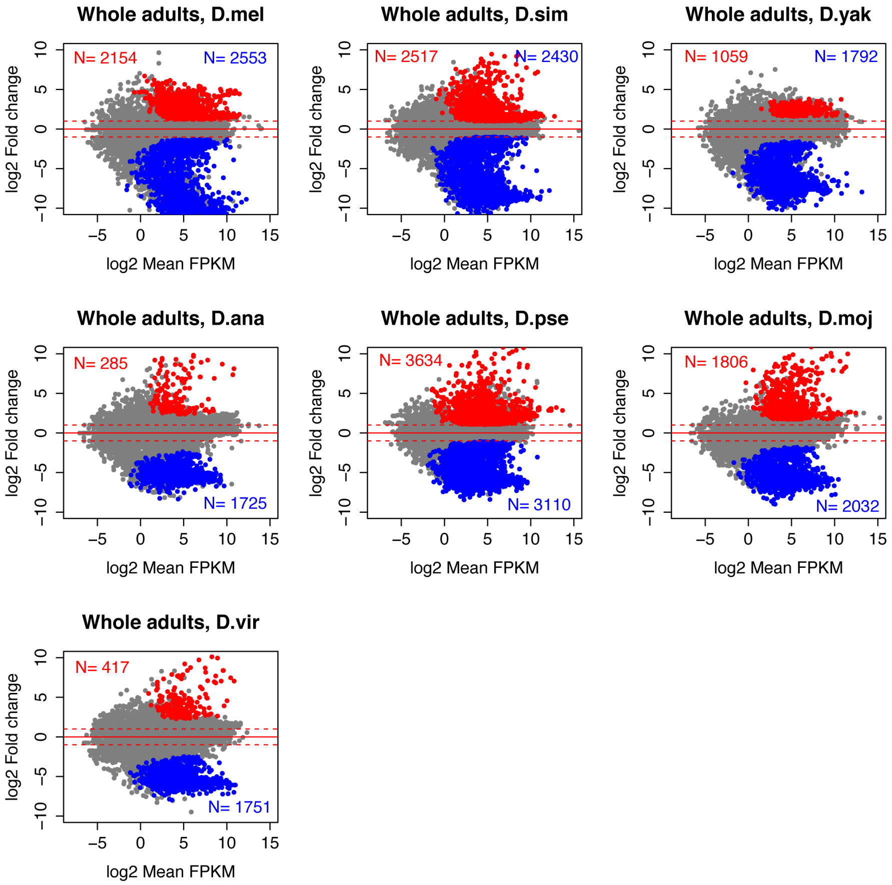


Figure 4.2: Sex-differential gene expression in whole adults of seven *Drosophila* species. Ratio vs intensity plots ("MA-plots") are shown for each species. Log2 ratio of female vs male expression is plotted on the y-axis, and average abundance is on the X-axis in units of FPKM. Species displayed (in order): *D. melanogaster, D. simulans, D. yakuba, D. ananassae, D. pseudoobscura, D. mojavensi*s, and *D. virilis*.

In whole adult tissue, sex-differential transcription is substantial, with 36% of genes sex-differentially expressed in *D. melanogaster*. As was seen in previous experiments with microarrays, the general pattern is for more male-biased expression than female-biased expression, with the notable exception of *D. pseudoobscura* (Zhang, Sturgill et al. 2007). In this species, complex evolutionary forces have shaped the transcriptome due to the fusion of an autosomal arm to the X chromosome, creating a "neo-X" chromosome (Richards, Liu et al. 2005; Sturgill, Zhang et al. 2007).

Even if gene expression between orthologs is equal, different species may generate sex-differential processed transcripts, by employing different post-transcriptional regulation. For example, a conserved cassette exon may be included in males specifically in one species and in females specifically in another, while overall gene expression is the same. To explore this possibility, we examined differences in sex-differential splicing between the species.

4.2.2 Interspecific splicing comparison

To compare splicing events between species, we took a *melanogaster*-centric approach where we used splicing event definitions generated in *D. melanogaster* and projected them onto the other species. Since each splicing event is defined by sets of junctions (Chapter 2), we can quantify homologous events by comparing coverage values among homologous junctions.

To expand our sampling of the *Drosophila* lineage, we obtained reference genome sequence and RNA-Seq data for whole adult males and females of eight additional species (Stephen Richards, personal communication; see Materials and methods). Despite the lack of defined orthologous genes in these species, we can assay splicing differences with junction coverages that are matched through whole genome alignments ("liftovers"). Since each splicing event is composed of mutually exclusive junctions, we require each path to be detected in either sex. This requirement lends confidence that the event is lifted over accurately between species.

We first compared splicing using the Proportion Spliced In "PSI" metric calculated by Spanki (Chapter 2). We compared the female vs male ΔPSI values for detected orthologous splicing events between *D. melanogaster* and 13 other Drosophila species, including two different strains of *D. simulans* (Figure 4.3). Since large differences in proportions may result from low abundance events, we required each event to have detected coverage >= 10 in both the "inclusion" and "exclusion" paths in either sex.

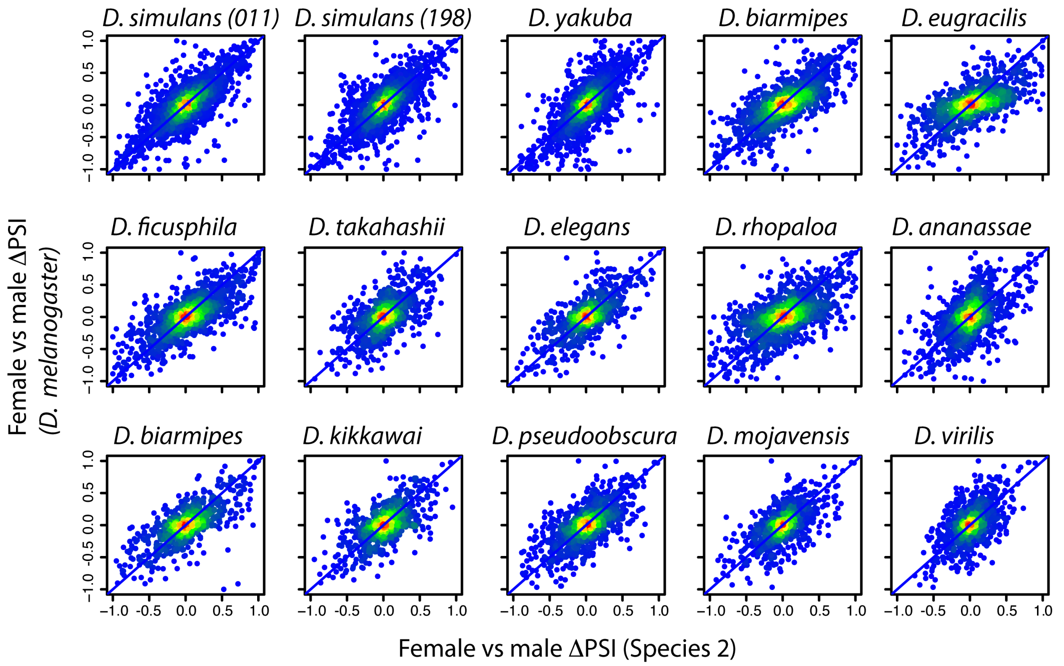


Figure 4.3: Differences in female vs male proportion spliced in (ΔPSI), between *D. melanogaster* and 14 other species of *Drosophila* and (including two strains of *D. simulans)*. Colors in each scatter plot represent density, with the greatest number of values in red. The blue represents the diagonal (not a calculated trend line).

Comparing events between species revealed ΔPSI to be generally conserved between species, with few diametric shifts in proportion (Figure 4.3). In each pairwise comparison, the greatest density of values centered at equal inclusion/exclusion proportions in each sex (ΔPSI close to zero), with the remainder of values along the diagonal, and few in the upper left and lower right quadrants (where dramatic differences would be).

Since we did not observe large stochasticity in sex-differential PSI between species, we asked where this metric may change along a lineage. This would provide evidence that a change is a heritable, adaptive change in splicing regulation, rather than random drift. To see where patterns in PSI might change along a lineage, we examined orthologous events across several species, rather than pairwise to *melanogaster* (Figure 4.4).

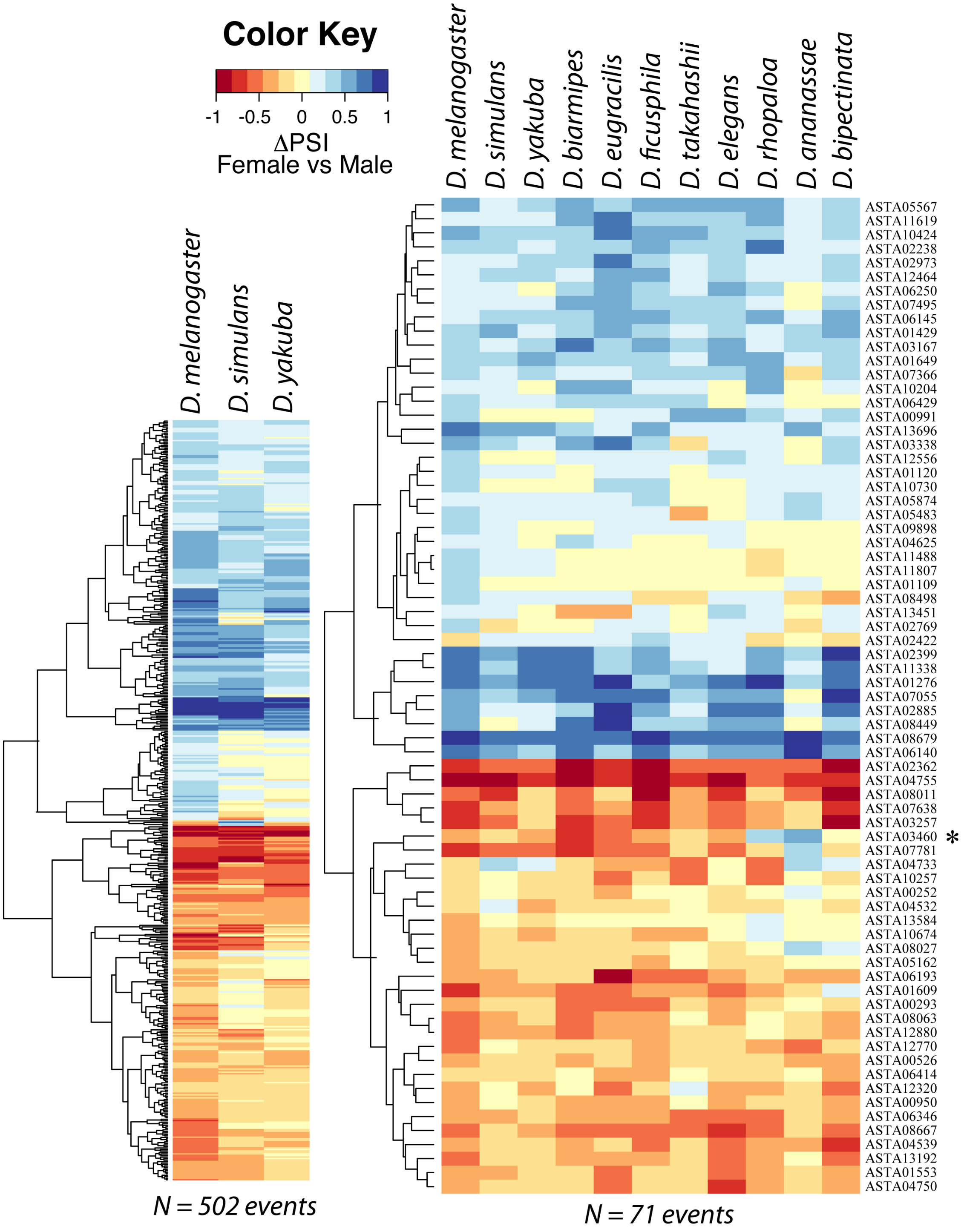


Figure 4.4: Heatmap of female vs male ΔPSI for 502 events detected in all species of the *melanogaste*r subgroup, and 71 events detected in 11 species of the melanogaster group. An asterisk notes one example event with a large change in PSI direction (ASTA03460) in the gene *Pfk*.

We extracted splicing events that showed a large sex difference in PSI in *D. melanogaster* whole adults (940 events with |ΔPSI| > 0.2, q-value < 0.05, and minimum 10 junction reads in each path). We then compared splicing events at two different phylogenetic resolutions, by compiling data for the same events that were confidently detected in each species of the *melanogaster* group and among the more closely related *melanogaster* subgroup (Figure 4.4). We clustered these values with hierarchical clustering, and looked for events that differed between lineages. We observed variation in degree of ΔPSI between species, but few cases of switching between inclusion or exclusion form. One example of a switch in ΔPSI is in the gene *Phosphofructokinase (Pfk)*, which is a component of the glycolytic pathway and that changes inclusion/exclusion preference in *D. ananassae.* This gene encodes an estrogen receptor domain that is conserved across Drosophila, and is essential for detecting signals for upregulation of carbohydrate metabolism in larval development (Tennessen, Baker et al. 2011).

4.2.3 Conservation of sex-determination components

It has been previously observed that *Sxl* dimorphic splicing is not conserved in all insects (Sanchez 2008; Cline, Dorsett et al. 2010). Taken with the fact that orthologs of *dsx* are conserved in diverse lineages from worms to human, leads to the hypothesis that the *tra*-*dsx* axis was the ancient sex-determination mechansism at the divergence of *Drosophil*a from the other Dipterans, and that *Sxl* was recruited later to serve as a master switch (Haag and Doty 2005). To extend these observations, we examined in the conservation of splicing of *dsx* and *Sxl* transcripts in other *Drosophila* species.

We quantified the *dsx* and *Sxl* sex differential splicing events based on lifted over junctions, and identified sex-differential regulation in each species (Figure 4.5). Since sequencing depth varied between species and between sex, we enforced an abundance cutoff of ten junction spanning reads in either inclusion or exclusion form in each sex. This analysis confirmed sex-specific splicing of *dsx* in eight non-melanogaster species, spanning the breadth of the Drosophila lineage (~40 million years (Clark, Eisen et al. 2007)). For *Sxl*, we could confirm sex-differential splicing in ten species within the *melanogaster* group (including the *ananassae* subgroup). In more distant species, the orthologous splicing event could not be identified using liftovers, which may be due to alignment difficulties arising from the rapid divergence of this locus (Cline, Dorsett et al. 2010). By manually curating alignments, identification of the orthologous junctions in more distant species was achieved, and sex-differential splicing was observed (Figure 4.5).

We also examined reference genomes of multiple species, so see if there were differences that could affect sex-differential splicing. We used fuzznuc (Ensembl) to count occurrences of the TRA binding site (TC[AT][AT]C[AG]ATCAACA) in each Drosophila species. We found ~100 occurrences in *D. melanogaster*, but substantially more in some species (~200 in *D. grimshawi* and *D. persimilis*) and less in others (~50 in *D. biarmipes*).

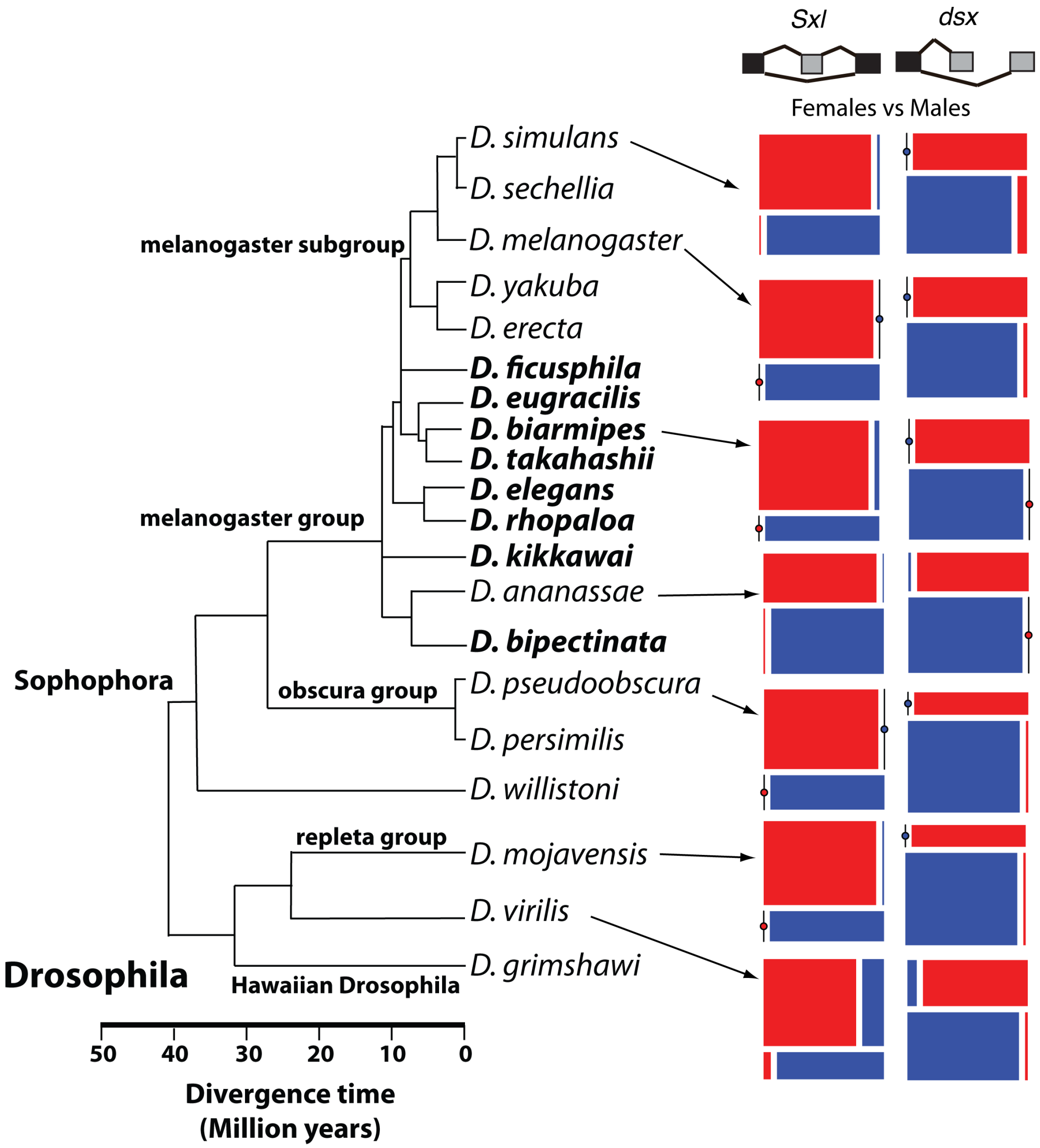


Figure 4.5 Conservation of *dsx* and sex-specific splicing in multiple *Drosophila* species. Data are shown for each species where detection was abundant in each sex (>= 10 reads in each sex, and in the inclusion and exclusion path of either sex).

Section 4.3 Discussion and conclusions

In whole adult samples, we identified sex-differential splicing in 940 events, and differential gene expression in 4707 genes in *D. melanogaster*. This suggests that transcriptional differences between sexes are greater than splicing differences in the whole animal. Comparing between species, we find that sex differential splicing of these events does not frequently change direction within the genus, but does vary by degree, as was shown previously with sex-biased transcription (Zhang, Sturgill et al. 2007).

We also confirm sex-differential splicing of *dsx* in eight non-melanogaster species. Coverage data from our RNA-Seq data suggest that *Sxl* orthologs in species distant from *D. melanogaster* have different exon/intron structure, which complicates the identification of an orthologous sex-differential splicing event. Nevertheless, sex-differential splicing was detectable in these species. Additional experiments at the protein level are required to better understand the role in sex-determination of *Sxl* in these species.

Section 4.4 Materials and methods

Annotation

Transcript models were obtained from ENSEMBL, and reference genome sequence was obtained from GenBank or the UCSC Genome Browser (Kent, Sugnet et al. 2002). Genome references used for each species were: *D. simulans* (droSim1), *D. yakuba* (droYak2), *D. ananassae* (droAna3), *D. mojavensis* (droMoj3) and *D. virilis* (droVir3) from UCSC; and *D. pseudoobscura* (Dpse_2.0), *D. biarmipes* (Dbia_1.0), *D. bipectinata* (Dbip_1.0), *D. elegans* (Dele_1.0), *D. eugracillis* (Deug_1.0), *D. ficusphila* (Dfic_1.0), *D. kikkawai* (Dkik_1.0), *D. rhopaloa* (Drho_1.0) and *D. takahashi* (Dtak_1.0). Annotation versions were *D. melanogaster* r5.39 (For ortholog analysis, supplemented with ModEncode v2 for splice junction analysis), *D. pseudoobscura* r2.22 (HGSC2.13 in ENSEMBL), *D. simulans* r1.3, D.

yakuba r1.3, *D. ananassa*e r1.3, *D. mojavensis* r1.3, and *D. virilis* r1.2. Orthology relationships were obtained from OrthoDB (Waterhouse, Zdobnov et al. 2011).

Sample collection and library construction

Whole adult sample collection for the *D. meanogaster, D. simulans, D. yakuba, D. ananassae , D. pseudoobscura, D. mojavensis*, and *D. virilis* samples is described in Zhang et al, 2007 (Zhang, Sturgill et al. 2007). Data for the additional 8 species were obtained from Baylor (S. Richards, personal communication), and are available in the short read archive (SRA). The additional eight species of Drosophila were chosen for genome sequencing and RNA-Seq analysis to support the modENCODE project. They were therefore chosen to fill "phylogenetic discovery gaps" in the Drosophila lineage (Piano and Cherbas 2010).

Most of these new species are in the *melanogaster* species group. These species are all South Asian. *D. ficusphila*, *D. biarmipes*, *D. eugracilis*, *D. takahashii*, and *D. rhopaloa* are fruit-feeding; while *D.elegans* is flower-feeding. *D. kikkawai* is the first sequenced member of the montium subgroup, which diverged from the *melanogaster* species group after it's divergence from the *ananassae* subgroup. It is originally from south Asia, but invasively widespread in Africa and South America. *D. bipectinat*a is related to *D. ananassae*, in the morphologically diverse ananassae lineage. These species have broadly divergent phenotypical characteristics, with makes them an emerging model from phenotypic evolution (Matsuda, Ng et al. 2009; Piano and Cherbas 2010).

Read alignment and liftovers

Read alignments were performed with Tophat v.2.0.4. Gene level estimates were obtained using Cufflinks (2.0.2). Junction quantifications were obtained using Spanki (Chapter 2), and event definitions are defined by AStalavista (Sammeth, Foissac et al. 2008).

We extracted coordinates of all splice junctions from *D. melanogaster* annotation, and generated BED files where each junction was represented by a 20bp anchor region in each adjoining exon. Liftovers were then performed with the reference genomes for each non-*melanogaster* species using the UCSC liftover tools (Kent, Sugnet et al. 2002). Chain files to perform the liftovers were made with *lastz* (an improved successor of *blastz*,(Harris 2007)).

We required that each lifted-over junction bordered donor/acceptor motifs (GT-AG, GC-AG, or AT-AC). This process obtained liftovers with valid donor/acceptor motifs for 63% (*D. mojavensis*) to 92% (*D. yakuba*) of junctions in the *D. melanogaster* reference annotation. The proportions of lifted-over junctions with major form vs minor form donor/acceptor motifs mirrored that of the reference annotation (~2 orders of magnitude more major form than minor form).

Section 4.5 Author contributions

I performed all computational analyses described in this chapter, with the exception of alignments for a subset of these data, which were performed by Zhen-Xia Chen. I am the primary author of the text. The majority of data used in this chapter was generated as part of the modENCODE project (http://www.modencode.org)

Biological samples were generated and RNA-Seq experiments were performed in the Oliver lab by Carlo Artieri, John H. Malone, Nico Mattiuzo and Harold Smith; and at Baylor university (Stephen Richards) and UC Davis (Artyom Kopp).

Chapter 5: Conclusions and outlook

Section 5.1 Introduction

Our results demonstrate the complexity of the transcriptome, and underscore the power of RNA-Seq to reveal this complexity. When microarrays were the dominant high-throughput technology, researchers became used to reporting gene-level fold changes as the primary analysis result. To take advantage of the additional resolution of RNA-Seq, it is now necessary to analyze alternative promoter use, relative isoform proportions, and splicing differences at the level of pairwise events and individual junctions. The Spanki toolkit is a big advance toward making these analyses possible, but technology continues to change rapidly, and software tools will have to continue to change to keep pace. Below I describe some outlook on the state of RNA-Seq technology and its future directions, and describe how this technology will provide gains in understanding the sex-determination network.

Section 5.2 RNA-Seq and analysis methods

5.2.1 Experimental / sequencing directions

The first sequencing instrument our lab had access to was an Illumina GAI (Illumina, San Diego, CA) and our first RNA-Seq experiments our lab performed generated 36bp single-end reads. These experiments produced 1-2 million unique mapped reads (See GEO entry GSE20348) per lane. More recent experiments (Such as the data described in Chapters 2 and 3) use data produced on an Illumina HiSeq

2000 instrument. These experiments yield about two orders of magnitude more usable reads per lane (> 100 million uniquely mapped paired-end reads).

With this great abundance of reads from one lane, it is now possible to multiplex and sequence multiple samples in one lane. This is made possible with indexed barcodes that are built in to Illumina adapter sequences (Wang, Si et al. 2011). Illumina kits are available to do 12 samples per lane (Illumina, San Diego, CA) but it is possible to do much more. Some users have reported successfully sequencing 96 samples per lane (Li, Schmieder et al. 2012).

Our own experiments (Chapter 2) and others have made the point the replication is very important to reliable inference from RNA-Seq data (Fang and Cui 2011). The expense of these experiments has lead many researchers to forgo biological replication, but multiplexing will allow greater replication to be possible at more reasonable cost. The surprising thing I learned from our experiments is that greater sequencing depth is not helpful to detect rare transcripts, as the vast majority of reads go to the most abundant transcripts, and read depths greater than 50 million reads detect more false positives than new true positives. These results reinforce the conclusion that greater replication at moderate depth is preferable to high depth of few replicates.

Despite the limitations of current sequencing technology, our simulations have shown that splice junction detection is highly accurate and sensitive (Chapter 2). However, lack of replication remains a problem because variance from biological replication is high, confounding estimates of between-sample differences.

Further down the road is the next quantum leap in sequencing technology; so called ‘third-generation sequencing’ (Schadt, Turner et al. 2010). In addition to longer read lengths, machines using these technologies also provide lower costs, and greater ease of use due to fewer reagents and lack of requirements for sophisticated optics. One machine using this new class of technology is from Ion Torrent (Life Technologies, Grand Island, NY), and has recently been put into use at the NIH Intramural Sequencing Center (NISC, Robert Blakesley, personal communication), but use of this technology is not yet widespread.

The impact of this changing technology on downstream RNA-Seq analysis is not clear. Longer read lengths that approach whole single-molecule sequencing will remove the need for transcript assembly and remove the ambiguity with mapping reads to transcripts. Without the need for library construction, there will no longer be the biases associated with fragmentation and amplification. It is a safe bet however that as with any new technology, new unanticipated biases will arise to replace old ones.

5.2.2 Future directions for Spanki

Software to analyze RNA-Seq data are still not mature, but a great deal of progress has been made, and some standards are beginning to emerge. Primary short read aligners perform well, and there are now several options that perform the task similarly (Garber, Grabherr et al. 2011; Grant, Farkas et al. 2011). Gene level abundance estimates and comparisons are calculated reliably by DESeq and Cufflinks (Anders and Huber 2010; Trapnell, Roberts et al. 2012).

Software to analyze splicing events and junctions are less well developed (Chapter 1 and Chapter 2), and this is where Spanki fills a major gap. I envisioned the toolkit to take a place along with other toolkits that have emerged as standards to handle particular data types, such as Samtools (Li, Handsaker et al. 2009) for alignment files and BEDTools (Quinlan and Hall 2010) for coordinate based quantitative data. Spanki was designed to perform a variety of analyses centered along one data type: a splice junction. Since Spanki is a modular Python package, new tasks can be easily added, along with wrappers for other junction-centered analyses.

A very useful addition to Spanki would be an analysis of intron sequence, using a rigorous model of donor and acceptor sites, branch points, and polypyrimidine tracts. This would serve two purposes: 1) provide a measure of splice site "strength" to compare with observed rates of inclusion or intron retention, and 2) serve as filtering criteria for false positive junctions. Support vector machines (SVMs) are in wide use for this task. I have built a utility that intersects junction coordinates and pre-computed SVM predictions (Sonnenburg, Schweikert et al. 2007), however these predictions are out of date. An alternative approach would be to use a Feature Generation Alogorithm (FGA), which has been trained and used successfully in Arabidopsis (Dogan, Getoor et al. 2007).

Section 5.3 Sex determination

5.3.1 An expanded sex-determination pathway

The major challenge to understanding sex-determination in the coming decade will likely be to fill in the vast gaps of knowledge about downstream components. The results I've described suggest that differences in splicing by sex are modest in Drosophila heads (Chapter 2), but these results provide strong evidence for several novel interesting targets of sex differences in regulation.

The gene *found in neurons* (*fne*) was detected as sex-differentially spliced in wildtype heads (Chapter 2). This regulatory event has also been validated by qPCR (Marie-Laure Samson, personal communication). FNE is a member of the embryonic-lethal abnormal vision (ELAV) gene family of RNA-binding proteins, which is conserved across diverse eukaryotic lineages (Samson and Chalvet 2003; Pascale, Amadio et al. 2008). Wildtype *fne* is required for robust male courtship behavior (Zanini, Jallon et al. 2012). FNE does not regulate splicing directly, but binds RNA and modulates transcript stability(Samson and Chalvet 2003), providing an interesting example of sex-differential post-transcriptional regulation through this mechanism. An important analysis that remains is to identify whether this splicing event is regulated by TRA, or some other splicing factor.

Rtnl1 is a membrane protein localized to the endoplasmic reticulum (ER) (Wakefield and Tear 2006) and has a role inter-male aggressive behavior (Edwards, Zwarts et al. 2009), olfactory response (Sambandan, Yamamoto et al. 2006), and motor axon development (O'Sullivan, Jahn et al. 2012). I observed sex differences in alternative promoter use in this gene (Chapter 2). I also observed a loss of this promoter difference in *Doa* mutants, suggesting that this promoter preference in regulated by trans-factors that are regulated by splicing.

This analysis also identified junction level sex differences in Drosophila heads that were not classified as events. We found that 12.5% of splice junctions were detected only in females, and 14% were detected only in males, and these were mainly low abundance. These junction detections passed our qualitative filtering criteria, so they are unlikely to be artifact, but their sex-specificity may be due to sequencing depth. To confidently compare rare transcripts between sexes, capture based methods to increase their abundance are required (Mercer, Gerhardt et al. 2012).

The next step in completing the picture of sex determination is to make the transition from the transcriptome to the proteome. An additional layer of regulation may be at work at the level of translation that may be important to sexual development. A novel application of next-generation sequencing technology is ribosomal profiling (Ingolia, Brar et al. 2012), which allows a quantitative analysis of translation in vivo.

It is clear that sex-differential transcriptome regulation is an interacting network of different types of regulation, which makes them difficult to identify individually. The task of providing evidence for phenotypes and genetic interactions of these targets will be a long process, but it will bring us closer to a complete picture of the global sex-determination network.

5.3.2 Splicing divergence in the Drosophila genus

The comparison of splicing between species is a new field with many challenges in analysis that have not been addressed in the literature (Romero, Ruvinsky et al. 2012). Our results provide a framework for comparing individual events between species using detected splice junctions (Chapter 4). We used this approach to compare events in 13 species of Drosophila, and perform additional comparisons to help understand the mechanisms that shaped transcriptome differences between these species.

A finer grained analysis of transcript model differences between species will require high quality transcript predictions in each species, informed by empirical evidence from ESTs and RNA-Seq. This is a difficult task, and current denovo assemblers such as Cufflinks (Trapnell, Williams et al. 2010) and Trinity (Grabherr, Haas et al. 2011) have not been able to adequately construct reliable models. Efforts are currently under way to apply NCBI's annotation pipeline to generate annotation in non-*melanogaster* Drosophila species (GNOMON, Terrence Murphy, personal communication).

The divergence of *Sxl* within the *Drosophila* lineage raises the possibility that there are diverse sex-determination pathways within the genus itself. For example, I found that there are nearly twice as many TRA binding sites in the *D. persimilis* and *D. grimshawi* reference genomes than in *D. melanogaster*, suggesting a possible expansion of TRA targets in those species.

Low abundance 'noisy' splicing has been hypothesized to be important in humans (Pickrell, Pai et al. 2010), and this may be important in Drosophila evolution as well. Many low-abundance splicing events were detected in head tissue, often occurring in one sex. I also observed aberrant splicing in *dsx* transcripts when *Doa* was disrupted by mutation. This mutational lability of *dsx* transcript splicing may be illustrative of rare splicing errors that can occur in nature that can then be available for selection. This leads to a hypothesis that rare splicing errors are a major source of novelty in the transcriptome.

Appendices

Appendix 1: Read quality analysis: A case study in RNA-Seq QC

Appendix 1: Read quality analysis - a case study

Confidence in the quality of the primary data is key to analysis of RNA-Seq data. Quality control is not always straightforward, as there are a variety of problems that can occur, from problems with reagents to Sequencing errors to sample contamination. This appendix describes a detailed analysis that was performed on quality issues with some RNA-Seq samples, to provide a record of this analysis, and to illustrate the scope that a thorough quality analysis may take. This rigorous analysis can serve as a general case study on quality analysis for RNA-Seq data, and is described below.

Independent mRNA samples were used to prepare libraries as a 2nd biological replicate, with the exception of the *Doa*^{HD} F (female) sample, which was re-run as a technical replicate. For five of these samples, the fraction of reads that uniquely mapped to the reference genome was < 50%. The sample with the fewest reads mapping was the *Doa*^{HD} F sample (Run 57, Lane 1), with 14.6% of reads uniquely mapped. To investigate the cause of this low mapping, I performed an analysis of read quality and additional alignments to other references, for the lowest mapping sample (Run 57 Lane 1, “R57L1”).

I first analyzed the reads with FastQC (Simon Andrews), to examine basic quality characteristics. These results showed that the three prime ends of reads had low quality.

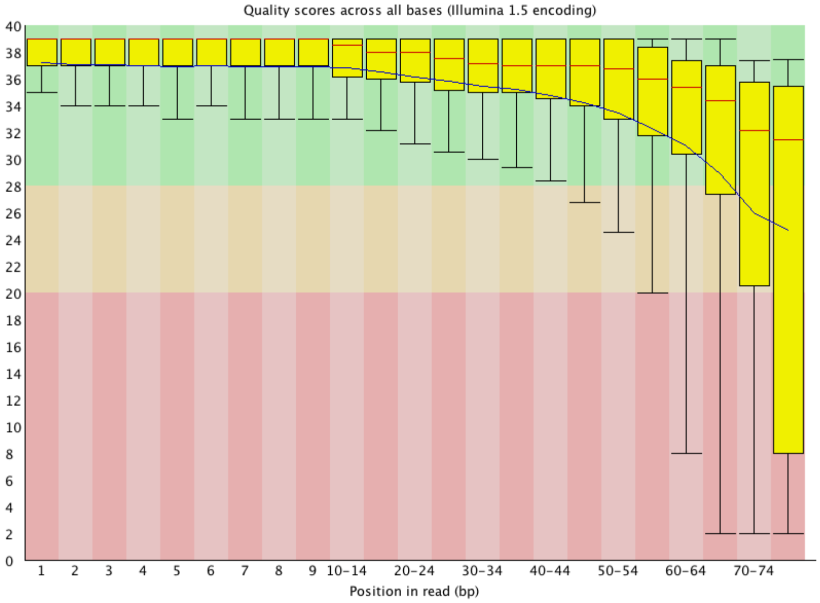


Figure A1.1 Per base quality scores for Run 57 Lane 1. Note that quality dips to below 28 at position 65.

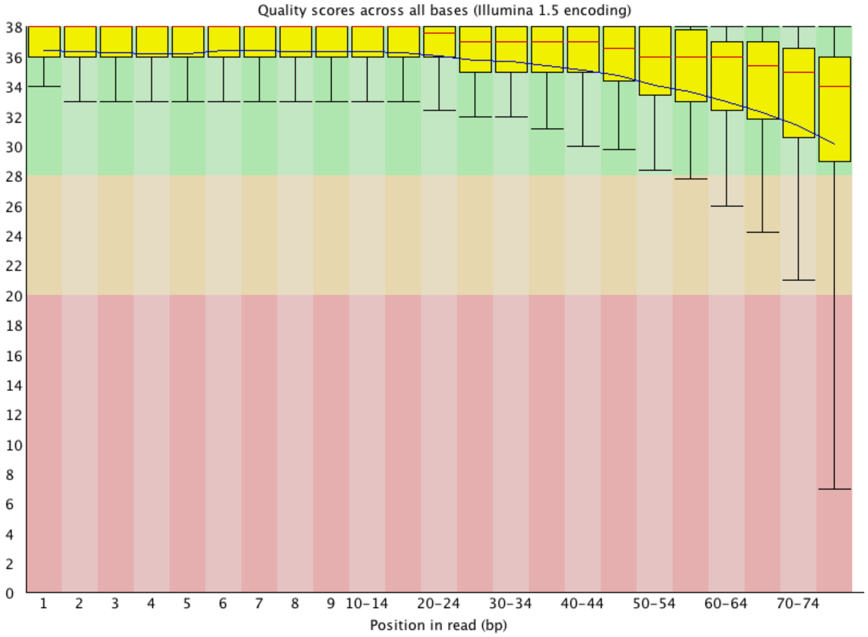


Figure A1.2 Per base quality scores for Run 50 Lane 1 (for comparison). Note that quality remains high throughout the read.

In addition to this low quality, a small amount (< 5% of total reads) was identified as adapter contamination ("Illumina Paired End PCR Primer 2" and "Illumina Paired End Adapter 2"). To determine whether there were other sources of contamination, I took reads that were unaligned after first-pass unique mapping with Tophat (Trapnell, Pachter et al. 2009), and performed assembly of these reads (Zerbino and Birney 2008) with default parameters. I selected the top 20 contigs by length and by abundance, and aligned them to the nr (non-redundant) database (NCBI) with BLASTN (Altschul, Gish et al. 1990).

This 2nd-pass mapping uncovered contaminating reads from Drosophila A virus isolate HD and Enterobacteria phage phiX174 (phi-X control). To assess the relative amounts of contamination, I created a new reference of these contaminants, along with the Illumina adapters and primers, and a set of transposable element Sequences. I also trimmed these reads to 50bp and re-aligned them to dm3 using an alternative version of Bowtie with reported greater sensitivity (Bowtie2, (Langmead, Trapnell et al. 2009))

From these results, I was able to estimate the various fractions of contamination and mappability problems for the set of reads in R57L1 (Table A1.1).

Reads uniquely aligned:	14.6%
Uniquely aligned at 50bp:	20.9%
Multi-mapped at 50bp:	14.8%
Adaptor/primer:	0.5%
Phi-X, virus:	2.5%
Ribosomal:	39.1%
Total explained:	***93.4%***

Table A1.1: Estimates of various alignment classes, as a percent of total reads.

These results show that the single largest source of unmapped reads derives from ribosomal protein genes. These reads are only mappable trimmed, and only with the alternative version of Bowtie (Bowtie2). The reason that they could not map with Tophat could not be determined. I also determined that trimming reads to 50bp can greatly increase the percent of reads mapping.

Following this analysis of run quality, the next step is to decide whether to use a different set of reference alignments for these reads. For example, I may trim the reads for the low mapping runs and obtain more mappings, but I may lose comparability with replicates mapped without trimming. To help make this decision, I performed correlation analysis for R57L1 and its replicate in an earlier run (R50L2).

I determined correlations for both raw read counts (HTSeq (Anders and Huber 2010), Figure A1.3), and for FPKMs (Cufflinks, (Trapnell, Roberts et al. 2012), Figure A1.4), for full length alignments and trimmed alignments. I found the highest correlation (0.96) for comparing two sets of full-length alignments together, using raw counts (Figure A1.3).

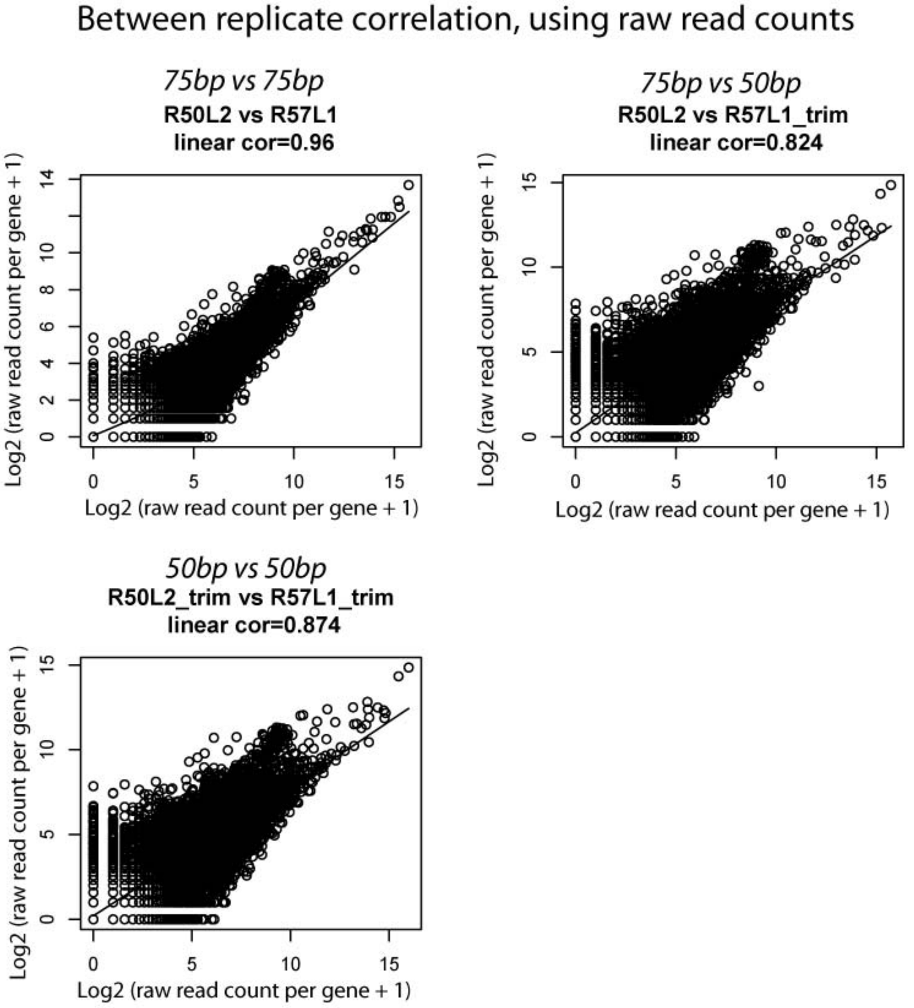


Figure A1.3 Between replicate correlations of 1million read subsets for *doa*^{HD} M sample, R50L2 (High % mapping) and R57L1 (Low percent mapping) Note that correlation is highest (0.96) when both sets are mapped the same way, full length. Correlation goes down when the poor quality sample is trimmed (0.82) and when both samples are trimmed (0.87).

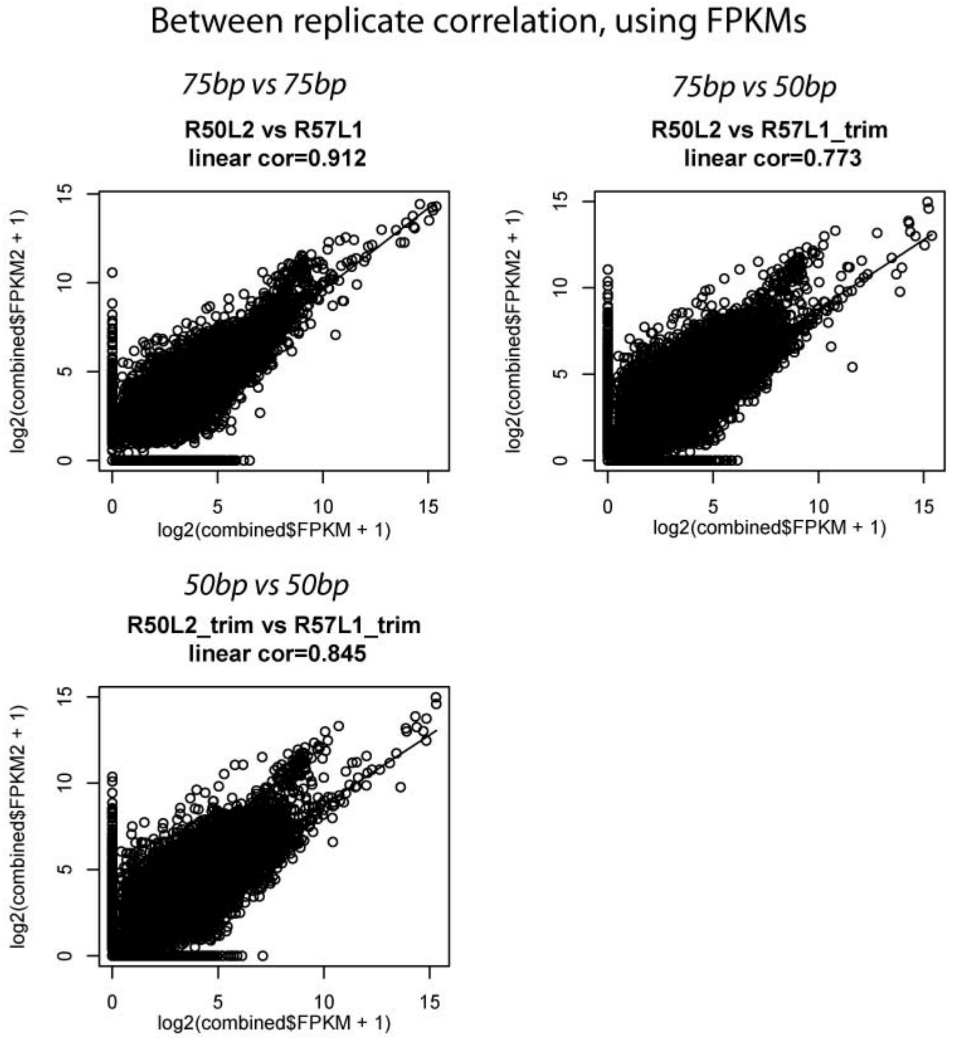


Figure A1.4 Between replicate correlation using gene level abundance estimates (in FPKM). Note that correlations are lower that in the comparison of raw read counts (Figure A1.3)

The between replicate correlation analysis showed that raw read counts within genes is better correlated between replicates than the probabilistic abundance estimates (FPKMs). It also showed that correlations were highest when using the same read length for each dataset, and when using the full length of the reads (Figure A1.3).

I also compared the percent of reads the map to the genome but do not overlap any annotated feature (identified by HTSeq as ‘no_feature’), and reads that map to genomic regions shared by multiple genes ('indentified by HTSeq as 'ambiguous') (Figure A1.5). Flybase release 5.44 was used as the reference in this analysis.

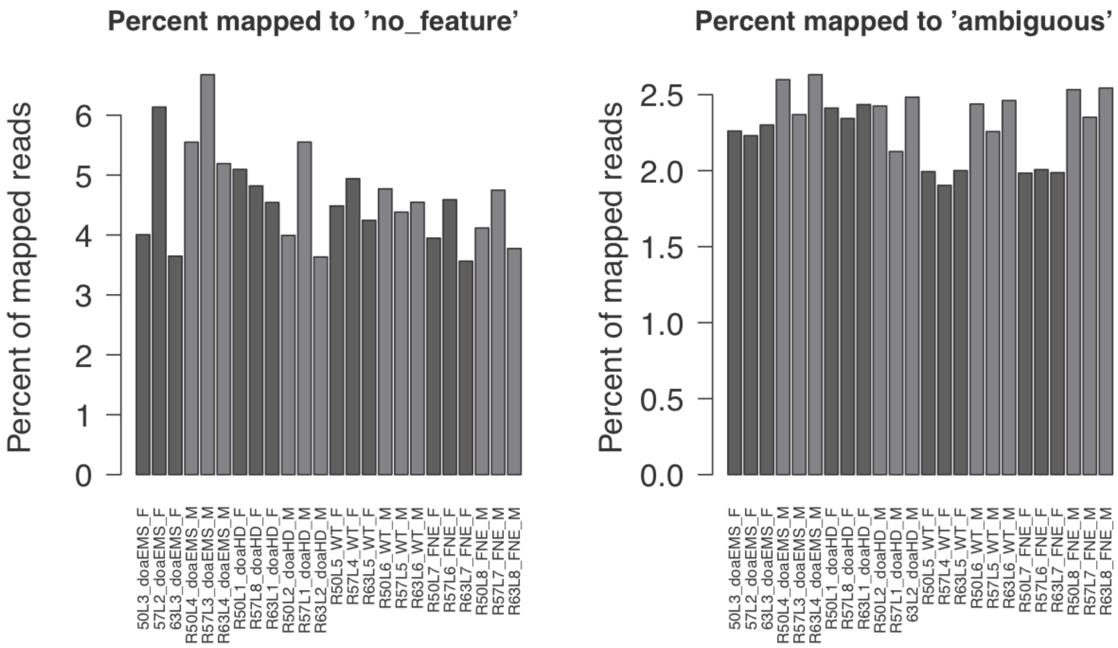


Figure A1.5 Comparison of the mapped reads that cannot be assigned to a gene feature, either due to no feature defined in the mapped location ('no_feature') or due to multiple overlapping features ('ambiguous'), for 24 samples. Columns are grouped by sample.

The low percent mapping in Run 57 is due mainly to transcripts from ribosomal protein genes. Increasing the percent mapping by trimming lowers the between-replicate correlations. Abundance estimates of non-ribosomal protein genes are replicable, when ribosomal protein genes with high abundance are excluded from normalization by Sequencing depth. For these reasons, the best approach is use full-length alignments for all samples.

Deciding how to preprocess reads and when to trim them requires careful analysis, as this case study illustrates. The proper course should be driven by the data, to maximize comparability between replicates primarily, and to preserve depth and resolution secondarily. Since RNA-Seq experiments are quite costly, few replicates are often generated. When a first pass alignment generates low percent mapping, this does not necessarily mean the data are not usable, and it is a good strategy to use all replicates available. However, all samples with low percent mapping should be examined to identify the cause, to identify potential systematic sources of error or contamination.

Bibliography

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J.Mol Biol. **215**(3): 403-410.

Ambrus, A. M., V. I. Rasheva, B. N. Nicolay and M. V. Frolov (2009). "Mosaic genetic screen for suppressors of the de2f1 mutant phenotype in Drosophila." Genetics **183**(1): 79-92.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Nature Proceedings.

Anders, S., A. Reyes and W. Huber (2012). "Detecting differential usage of exons from RNA-seq data." Genome Res.

Bashaw, G. J. and B. S. Baker (1997). "The regulation of the Drosophila msl-2 gene reveals a function for Sex-lethal in translational control." Cell **89**(5): 789-798.

BDGP. (2006). "BDGP Drosophila genome release 5." from http://www.fruitfly.org/sequence/release5genomic.shtml.

Beckmann, K., M. Grskovic, F. Gebauer and M. W. Hentze (2005). "A dual inhibitory mechanism restricts msl-2 mRNA translation for dosage compensation in Drosophila." Cell **122**(4): 529-540.

Bell, L. R., J. I. Horabin, P. Schedl and T. W. Cline (1991). "Positive autoregulation of sex-lethal by alternative splicing maintains the female determined state in Drosophila." Cell **65**(2): 229-239.

Bellen, H. J., C. Tong and H. Tsuda (2010). "100 years of Drosophila research and its impact on vertebrate neuroscience: a history lesson for the future." Nat Rev Neurosci **11**(7): 514-522.

Billeter, J. C., S. F. Goodwin and K. M. O'Dell (2002). "Genes mediating sex-specific behaviors in Drosophila." Adv Genet **47**: 87-116.

Black, D. L. (2003). "Mechanisms of alternative pre-messenger RNA splicing." Annu Rev Biochem **72**: 291-336.

Blencowe, B. J., S. Ahmad and L. J. Lee (2009). "Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes." Genes Dev **23**(12): 1379-1386.

Boue, S., I. Letunic and P. Bork (2003). "Alternative splicing and evolution." Bioessays **25**(11): 1031-1034.

Brooks, A. N., L. Yang, M. O. Duff, K. D. Hansen, J. W. Park, S. Dudoit, S. E. Brenner and B. R. Graveley (2011). "Conservation of an RNA regulatory map between Drosophila and mammals." Genome research **21**: 193-202.

Burtis, K. C. and B. S. Baker (1989). "Drosophila doublesex gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides." Cell **56**(6): 997-1010.

Celniker, S. E. and G. M. Rubin (2003). "The Drosophila melanogaster genome." Annu Rev Genomics Hum Genet **4**: 89-117.

Chang, P. L., J. P. Dunham, S. V. Nuzhdin and M. N. Arbeitman (2011). "Somatic sex-specific transcriptome differences in Drosophila revealed by whole transcriptome sequencing." BMC genomics **12**: 364.

Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver, T. A. Markow, T. C. Kaufman, M. Kellis, W. Gelbart, V. N. Iyer, D. A. Pollard, T. B. Sackton, A. M. Larracuente, N. D. Singh, J. P. Abad, D. N. Abt, B. Adryan, M. Aguade, H. Akashi, W. W. Anderson, et al. (2007). "Evolution of genes and genomes on the Drosophila phylogeny." Nature **450**(7167): 203-218.

Cline, T. W. (1984). "Autoregulatory functioning of a Drosophila gene product that establish es and maintains the sexually determined state." Genetics **107**(2): 231-277.

Cline, T. W., M. Dorsett, S. Sun, M. M. Harrison, J. Dines, L. Sefton and L. Megna (2010). "Evolution of the Drosophila feminizing switch gene Sex-lethal." Genetics **186**(4): 1321-1336.

Clutton-Brock, T. (2007). "Sexual selection in males and females." Science **318**(5858): 1882-1885.

Cuperlovic-Culf, M., N. Belacel, A. S. Culf and R. J. Ouellette (2006). "Microarray analysis of alternative splicing." OMICS **10**(3): 344-357.

Darwin, C. (1871). The Descent of Man, and Selection in Relation to Sex, J. Murray.

Davuluri, R. V., Y. Suzuki, S. Sugano, C. Plass and T. H. M. Huang (2008). "The functional consequences of alternative promoter use in mammalian genomes." Trends in Genetics **24**(4): 167-177.

De Bona, F., S. Ossowski, K. Schneeberger and G. Rätsch (2008). "Optimal spliced alignments of short sequence reads." Bioinformatics (Oxford, England) **24**: i174-180.

Demir, E. and B. J. Dickson (2005). "fruitless splicing specifies male courtship behavior in Drosophila." Cell **121**(5): 785-794.

Deng, X., J. B. Hiatt, D. K. Nguyen, S. Ercan, D. Sturgill, L. W. Hillier, F. Schlesinger, C. A. Davis, V. J. Reinke, T. R. Gingeras, J. Shendure, R. H. Waterston, B. Oliver, J. D. Lieb and C. M. Disteche (2011). "Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster." Nat Genet **43**(12): 1179-1185.

Dimon, M. T., K. Sorber and J. L. DeRisi (2010). "HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data." PloS one **5**: e13875.

Dimova, D. K., O. Stevaux, M. V. Frolov and N. J. Dyson (2003). "Cell cycle-dependent and cell cycle-independent control of transcription by the Drosophila E2F/RB pathway." Genes Dev **17**(18): 2308-2320.

Dogan, R. I., L. Getoor, W. J. Wilbur and S. M. Mount (2007). "SplicePort--an interactive splice-site analysis tool." Nucleic Acids Res **35**(Web Server issue): W285-291.

Edwards, A. C., S. M. Rollmann, T. J. Morgan and T. F. C. Mackay (2006). "Quantitative genomics of aggressive behavior in Drosophila melanogaster." PLoS genetics **2**: e154.

Edwards, A. C., L. Zwarts, A. Yamamoto, P. Callaerts and T. F. Mackay (2009). "Mutations in many genes affect aggressive behavior in Drosophila melanogaster." BMC Biol **7**: 29.

Ellegren, H. and J. Parsch (2007). "The evolution of sex-biased genes and sex-biased gene expression." Nat Rev Genet **8**(9): 689-698.

Fang, Z. and X. Cui (2011). "Design and validation issues in RNA-seq experiments." Brief Bioinform **12**(3): 280-287.

Filichkin, S. A., H. D. Priest, S. A. Givan, R. Shen, D. W. Bryant, S. E. Fox, W. K. Wong and T. C. Mockler (2010). "Genome-wide mapping of alternative splicing in Arabidopsis thaliana." Genome Research **20**(1): 45-58.

FitzGerald, P. C., D. Sturgill, A. Shyakhtenko, B. Oliver and C. Vinson (2006). "Comparative genomics of Drosophila and human core promoters." Genome Biol **7**(7): R53.

Fujii, S. and H. Amrein (2002). "Genes expressed in the Drosophila head reveal a role for fat cells in sex-specific physiology." The EMBO journal **21**: 5353-5363.

Gailey, D. A., J. C. Billeter, J. H. Liu, F. Bauzon, J. B. Allendorfer and S. F. Goodwin (2006). "Functional conservation of the fruitless male sex-determination gene across 250 Myr of insect evolution." Mol Biol Evol **23**(3): 633-643.

Garber, M., M. G. Grabherr, M. Guttman and C. Trapnell (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq." Nature Methods **8**: 469-477.

Garcia-Blanco, M. A., A. P. Baraniak and E. L. Lasda (2004). "Alternative splicing in disease and therapy." Nat Biotechnol **22**(5): 535-546.

Glazko, G. V. and M. Nei (2003). "Estimation of divergence times for major lineages of primate species." Mol Biol Evol **20**(3): 424-434.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol **29**(7): 644-652.

Grant, G. R., M. H. Farkas, A. Pizarro, N. Lahens, J. Schug, B. Brunk, C. J. Stoeckert, J. B. Hogenesch and E. A. Pierce (2011). "Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM)." Bioinformatics (Oxford, England).

Graveley, B. R., A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. van Baren, N. Boley, B. W. Booth, J. B. Brown, L. Cherbas, C. A. Davis, A. Dobin, R. Li, W. Lin, J. H. Malone, N. R. Mattiuzzo, D. Miller, D. Sturgill, et al. (2011). "The developmental transcriptome of Drosophila melanogaster." Nature **471**: 473-479.

Haag, E. S. and A. V. Doty (2005). "Sex determination across evolution: connecting the dots." PLoS Biol **3**(1): e21.

Hall, S. L. and R. A. Padgett (1996). "Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns." Science **271**(5256): 1716-1718.

Hanrahan, C. J., M. J. Palladino, B. Ganetzky and R. A. Reenan (2000). "RNA editing of the Drosophila para Na(+) channel transcript. Evolutionary conservation and developmental regulation." Genetics **155**(3): 1149-1160.

Harrington, E. D. and P. Bork (2008). "Sircah: a tool for the detection and visualization of alternative transcripts." Bioinformatics (Oxford, England) **24**: 1959-1960.

Harris, R. (2007). Improved pairwise alignment of genomic DNA. Ph.D., The Pennsylvania State University.

Hurst, L. D. (2002). "The Ka/Ks ratio: diagnosing the form of sequence evolution." Trends Genet **18**(9): 486.

Ingolia, N. T., G. A. Brar, S. Rouskin, A. M. McGeachy and J. S. Weissman (2012). "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments." Nat Protoc **7**(8): 1534-1550.

Innocenti, P. and E. H. Morrow (2010). "The sexually antagonistic genes of Drosophila melanogaster." PLoS Biol **8**(3): e1000335.

Jiang, L., F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras and B. Oliver (2011). "Synthetic spike-in standards for RNA-seq experiments." Genome Research.

Johnston, R. J., Jr., Y. Otake, P. Sood, N. Vogt, R. Behnia, D. Vasiliauskas, E. McDonald, B. Xie, S. Koenig, R. Wolf, T. Cook, B. Gebelein, E. Kussell, H. Nakagoshi and C. Desplan (2011). "Interlocked feedforward loops control cell-type-specific Rhodopsin expression in the Drosophila eye." Cell **145**(6): 956-968.

Joly, D., A. Korol and E. Nevo (2004). "Sperm size evolution in Drosophila: inter- and intraspecific analysis." Genetica **120**(1-3): 233-244.

Katz, Y., E. T. Wang, E. M. Airoldi and C. B. Burge (2010). "Analysis and design of RNA sequencing experiments for identifying isoform regulation." Nature methods **7**: 1009-1015.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler (2002). "The human genome browser at UCSC." Genome research **12**: 996-1006.

Keren, H., G. Lev-Maor and G. Ast (2010). "Alternative splicing and evolution: diversification, exon definition and function." Nat Rev Genet **11**(5): 345-355.

Kondrashov, F. A. and E. V. Koonin (2001). "Origin of alternative splicing by tandem exon duplication." Hum Mol Genet **10**(23): 2661-2669.

Kopelman, N. M., D. Lancet and I. Yanai (2005). "Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms." Nat Genet **37**(6): 588-589.

Kopp, A. (2006). "Basal relationships in the Drosophila melanogaster species group." Mol Phylogenet Evol **39**(3): 787-798.

Koscielny, G., V. Le Texier, C. Gopalakrishnan, V. Kumanduri, J.-J. Riethoven, F. Nardone, E. Stanley, C. Fallsehr, O. Hofmann, M. Kull, E. Harrington, S. Boué, E. Eyras, M. Plass, F. Lopez, W. Ritchie, V. Moucadel, T. Ara, H. Pospisil, A. Herrmann, et al. (2009). "ASTD: The Alternative Splicing and Transcript Diversity database." Genomics **93**: 213-220.

Kpebe, A. and L. Rabinow (2008). "Alternative promoter usage generates multiple evolutionarily conserved isoforms of Drosophila DOA kinase." Genesis **46**(3): 132-143.

Labaj, P. P., G. G. Leparc, B. E. Linggi, L. M. Markillie, H. S. Wiley and D. P. Kreil (2011). "Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling." Bioinformatics **27**(13): i383-391.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Lee, T. I. and R. A. Young (2000). "Transcription of eukaryotic protein-coding genes." Annu Rev Genet **34**: 77-137.

Li, B., V. Ruotti, R. M. Stewart, J. A. Thomson and C. N. Dewey (2010). "RNA-Seq gene expression estimation with read mapping uncertainty." Bioinformatics (Oxford, England) **26**: 493-500.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics (Oxford, England) **25**: 2078-2079.

Li, J. W., R. Schmieder, R. M. Ward, J. Delenick, E. C. Olivares and D. Mittelman (2012). "SEQanswers: an open access community for collaboratively decoding genomes." Bioinformatics **28**(9): 1272-1273.

Li, Q., J.-A. Lee and D. L. Black (2007). "Neuronal regulation of alternative pre-mRNA splicing." Nature reviews. Neuroscience **8**: 819-831.

Luo, S. D., G. W. Shi and B. S. Baker (2011). "Direct targets of the D. melanogaster DSXF protein and the evolution of sexual development." Development **138**(13): 2761-2771.

Lynch, K. W. and T. Maniatis (1996). "Assembly of specific SR protein complexes on distinct regulatory elements of the Drosophila doublesex splicing enhancer." Genes Dev **10**(16): 2089-2101.

Malko, D. B., V. J. Makeev, A. A. Mironov and M. S. Gelfand (2006). "Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes." Genome Res **16**(4): 505-509.

Malone, J. H. and B. Oliver (2011). "Microarrays, deep sequencing and the true measure of the transcriptome." BMC Biol **9**: 34.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Res **18**(9): 1509-1517.

Martin, J. A. and Z. Wang (2011). "Next-generation transcriptome assembly." Nature Reviews Genetics **12**: 671-682.

Matson, C. K. and D. Zarkower (2012). "Sex and the singular DM domain: insights into sexual regulation, evolution and plasticity." Nat Rev Genet **13**(3): 163-174.

Matsuda, M., C. S. Ng, M. Doi, A. Kopp and Y. N. Tobari (2009). "Evolution in the Drosophila ananassae species subgroup." Fly (Austin) **3**(2): 157-169.

McIntyre, L. M., L. M. Bono, A. Genissel, R. Westerman, D. Junk, M. Telonis-Scott, L. Harshman, M. L. Wayne, A. Kopp and S. V. Nuzhdin (2006). "Sex-specific expression of alternative transcripts in Drosophila." Genome Biol **7**(8): R79.

McQuilton, P., S. E. St Pierre and J. Thurmond (2012). "FlyBase 101--the basics of navigating FlyBase." Nucleic acids research **40**: D706-714.

Meise, M., D. Hilfiker-Kleiner, A. Dubendorfer, C. Brunner, R. Nothiger and D. Bopp (1998). "Sex-lethal, the master sex-determining gene in Drosophila, is not sex-specifically regulated in Musca domestica." Development **125**(8): 1487-1494.

Mercer, T. R., D. J. Gerhardt, M. E. Dinger, J. Crawford, C. Trapnell, J. A. Jeddeloh, J. S. Mattick and J. L. Rinn (2012). "Targeted RNA sequencing reveals the deep complexity of the human transcriptome." Nat Biotechnol **30**(1): 99-104.

Meyer, D. and K. Hornik (2006). "The Strucplot Framework : Visualizing Multi-way Contingency Tables with vcd." October **17**.

Meyer, F. and B. Moussian (2009). "Drosophila multiplexin (Dmp) modulates motor axon pathfinding accuracy." Dev Growth Differ **51**(5): 483-498.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon (2002). "Network motifs: simple building blocks of complex networks." Science **298**(5594): 824-827.

Modrek, B. and C. J. Lee (2003). "Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss." Nat Genet **34**(2): 177-180.

Moehring, A. J., J. Li, M. D. Schug, S. G. Smith, M. deAngelis, T. F. Mackay and J. A. Coyne (2004). "Quantitative trait loci for sexual isolation between Drosophila simulans and D. mauritiana." Genetics **167**(3): 1265-1274.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Mullon, C., A. Pomiankowski and M. Reuter (2012). "Molecular evolution of Drosophila Sex-lethal and related sex determining genes." BMC Evol Biol **12**: 5.

Murphy, M. W., D. Zarkower and V. J. Bardwell (2007). "Vertebrate DM domain proteins bind similar DNA sequences and can heterodimerize on DNA." BMC Mol Biol **8**: 58.

Nagoshi, R. N., M. McKeown, K. C. Burtis, J. M. Belote and B. S. Baker (1988). "The control of alternative splicing at genes regulating sexual differentiation in D. melanogaster." Cell **53**: 229-236.

O'Sullivan, N. C., T. R. Jahn, E. Reid and C. J. O'Kane (2012). "Reticulon-like-1, the Drosophila orthologue of the Hereditary Spastic Paraplegia gene reticulon 2, is required for organization of endoplasmic reticulum and of distal motor axons." Hum Mol Genet **21**(15): 3356-3365.

Oshlack, A., M. D. Robinson and M. D. Young (2010). "From RNA-seq reads to differential expression results." Genome biology **11**: 220.

Pan, Q., O. Shai, L. J. Lee, B. J. Frey and B. J. Blencowe (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." Nature genetics **40**: 1413-1415.

Pascale, A., M. Amadio and A. Quattrone (2008). "Defining a neuron: neuronal ELAV proteins." Cell Mol Life Sci **65**(1): 128-140.

Piano, F. and P. Cherbas (2010). A proposal for comparative genomics in support of the modENCODE project.

Pickrell, J. K., A. A. Pai, Y. Gilad and J. K. Pritchard (2010). "Noisy splicing drives mRNA isoform diversity in human cells." PLoS genetics **6**: e1001236.

Pitnick, S., G. S. Spicer and T. A. Markow (1995). "How long is a giant sperm?" Nature **375**(6527): 109.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Rabinow, L. and M.-L. Samson (2010). "The role of the Drosophila LAMMER protein kinase DOA in somatic sex determination." Journal of genetics **89**: 271-277.

Rasheva, V. I., D. Knight, P. Bozko, K. Marsh and M. V. Frolov (2006). "Specific role of the SR protein splicing factor B52 in cell cycle control in Drosophila." Mol Cell Biol **26**(9): 3468-3477.

Richards, S., Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M. J. Hubisz, R. Chen, R. P. Meisel, O. Couronne, S. Hua, M. A. Smith, P. Zhang, J. Liu, H. J. Bussemaker, M. F. van Batenburg, S. L. Howells, S. E. Scherer, E. Sodergren, et al. (2005). "Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution." Genome Res. **15**(1): 1-18.

Roberts, A., C. Trapnell, J. Donaghey, J. L. Rinn and L. Pachter (2011). "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology **12**: R22.

Rogers, M. F., J. Thomas, A. S. Reddy and A. Ben-Hur (2012). "SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data." Genome biology **13**: R4.

Romero, I. G., I. Ruvinsky and Y. Gilad (2012). "Comparative studies of gene expression and the evolution of gene regulation." Nat Rev Genet **13**(7): 505-516.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods in molecular biology (Clifton, N.J.) **132**: 365-386.

Saccone, G., I. Peluso, D. Artiaco, E. Giordano, D. Bopp and L. C. Polito (1998). "The Ceratitis capitata homologue of the Drosophila sex-determining gene sex-lethal is structurally conserved, but not sex-specifically regulated." Development **125**(8): 1495-1500.

Salvemini, M., U. Mauro, F. Lombardo, A. Milano, V. Zazzaro, B. Arcà, L. C. Polito and G. Saccone (2011). "Genomic organization and splicing evolution of the doublesex gene, a Drosophila regulator of sexual differentiation, in the dengue and yellow fever mosquito Aedes aegypti." BMC evolutionary biology **11**: 41.

Salz, H. K. and J. W. Erickson (2010). "Sex determination in Drosophila: The view from the top." Fly **4**: 60-70.

Sambandan, D., A. Yamamoto, J. J. Fanara, T. F. Mackay and R. R. Anholt (2006). "Dynamic genetic interactions determine odor-guided behavior in Drosophila melanogaster." Genetics **174**(3): 1349-1363.

Sammeth, M., S. Foissac and R. Guigó (2008). "A general definition and nomenclature for alternative splicing events." PLoS computational biology **4**: e1000147.

Samson, M.-L. and F. Chalvet (2003). "found in neurons, a third member of the Drosophila elav gene family, encodes a neuronal protein and interacts with elav." Mechanisms of development **120**: 373-383.

Sanchez, L. (2008). "Sex-determining mechanisms in insects." Int J Dev Biol **52**(7): 837-856.

Sanchez, L., B. Granadino and M. Torres (1994). "Sex determination in Drosophila melanogaster: X-linked genes involved in the initial step of sex-lethal activation." Dev Genet **15**(3): 251-264.

Schadt, E. E., S. Turner and A. Kasarskis (2010). "A window into third-generation sequencing." Hum Mol Genet **19**(R2): R227-240.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Shearman, D. C. (2002). "The evolution of sex determination systems in dipteran insects other than Drosophila." Genetica **116**(1): 25-43.

Shen-Orr, S. S., R. Milo, S. Mangan and U. Alon (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." Nat Genet **31**(1): 64-68.

Shukla, J. N. and J. Nagaraju (2010). "Doublesex: a conserved downstream gene controlled by diverse upstream regulators." Journal of genetics **89**: 341-356.

Singer, G. A. C., J. J. Wu, P. Yan, C. Plass, T. H. M. Huang and R. V. Davuluri (2008). "Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array." BMC Genomics **9**.

Siwicki, K. K. and E. A. Kravitz (2009). "Fruitless, doublesex and the genetics of social behavior in Drosophila melanogaster." Curr Opin Neurobiol **19**(2): 200-206.

Sonnenburg, S., G. Schweikert, P. Philips, J. Behr and G. Rätsch (2007). "Accurate splice site prediction using support vector machines." BMC bioinformatics **8 Suppl 10**: S7.

Sorek, R. and G. Ast (2003). "Intronic sequences flanking alternatively spliced exons are conserved between human and mouse." Genome Res **13**(7): 1631-1637.

Sorge, S., N. Ha, M. Polychronidou, J. Friedrich, D. Bezdan, P. Kaspar, M. H. Schaefer, S. Ossowski, S. R. Henz, J. Mundorf, J. Ratzer, F. Papagiannouli and I. Lohmann (2012). "The cis-regulatory code of Hox function in Drosophila." EMBO J **31**(15): 3323-3333.

Sosnowski, B. A., J. M. Belote and M. McKeown (1989). "Sex-specific alternative splicing of RNA from the transformer gene results from sequence-dependent splice site blockage." Cell **58**(3): 449-459.

Sturgill, D., Y. Zhang, M. Parisi and B. Oliver (2007). "Demasculinization of X chromosomes in the Drosophila genus." Nature **450**(7167): 238-241.

Sugnet, C. W., K. Srinivasan, T. A. Clark, G. O'Brien, M. S. Cline, H. Wang, A. Williams, D. Kulp, J. E. Blume, D. Haussler and M. Ares, Jr. (2006). "Unusual intron conservation near tissue-regulated exons found by splicing microarrays." PLoS Comput Biol **2**(1): e4.

Takahashi, H., S. Kato, M. Murata and P. Carninci (2012). "CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks." Methods in molecular biology (Clifton, N.J.) **786**: 181-200.

Tarazona, S., F. García-Alcalde, J. Dopazo, A. Ferrer and A. Conesa (2011). "Differential expression in RNA-seq: a matter of depth." Genome research **21**: 2213-2223.

Telonis-Scott, M., A. Kopp, M. L. Wayne, S. V. Nuzhdin and L. M. McIntyre (2009). "Sex-specific splicing in Drosophila: widespread occurrence, tissue specificity and evolutionary conservation." Genetics **181**(2): 421-434.

Tennessen, J. M., K. D. Baker, G. Lam, J. Evans and C. S. Thummel (2011). "The Drosophila estrogen-related receptor directs a metabolic switch that supports developmental growth." Cell Metab **13**(2): 139-148.

Trapnell, C., L. Pachter and S. L. Salzberg (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nature Protocols **7**: 562-578.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-515.

Traut, W., T. Niimi, K. Ikeo and K. Sahara (2006). "Phylogeny of the sex-determining gene Sex-lethal in insects." Genome **49**(3): 254-262.

Venables, J. P., R. Klinck, C. Koh, J. Gervais-Bird, A. Bramard, L. Inkel, M. Durand, S. Couture, U. Froehlich, E. Lapointe, J.-F. Lucier, P. Thibault, C. Rancourt, K. Tremblay, P. Prinos, B. Chabot and S. A. Elela (2009). "Cancer-associated regulation of alternative splicing." Nature structural & molecular biology **16**: 670-676.

Venables, J. P., J. Tazi and F. Juge (2011). "Regulated functional alternative splicing in Drosophila." Nucleic acids research.

Verhulst, E. C., L. van de Zande and L. W. Beukeboom (2010). "Insect sex determination: it all evolves around transformer." Curr Opin Genet Dev **20**(4): 376-383.

Vosshall, L. B. (2007). "Into the mind of a fly." Nature **450**(7167): 193-197.

Wakefield, S. and G. Tear (2006). "The Drosophila reticulon, Rtnl-1, has multiple differentially expressed isoforms that are associated with a sub-compartment of the endoplasmic reticulum." Cell Mol Life Sci **63**(17): 2027-2038.

Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge (2008). "Alternative isoform regulation in human tissue transcriptomes." Nature **456**: 470-476.

Wang, L., Y. Si, L. K. Dedow, Y. Shao, P. Liu and T. P. Brutnell (2011). "A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq." PLoS ONE **6**(10): e26426.

Wang, L., Y. Xi, J. Yu, L. Dong, L. Yen and W. Li (2010). "A statistical method for the detection of alternative splicing using RNA-seq." PloS one **5**: e8529.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature reviews. Genetics **10**: 57-63.

Waterhouse, R. M., E. M. Zdobnov, F. Tegenfeldt, J. Li and E. V. Kriventseva (2011). "OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011." Nucleic Acids Res **39**(Database issue): D283-288.

Williams, T. M. and S. B. Carroll (2009). "Genetic and molecular insights into the development and evolution of sexual dimorphism." Nat Rev Genet **10**(11): 797-804.

Xing, Y. and C. Lee (2006). "Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes." Nat Rev Genet **7**(7): 499-509.

Yamamoto, A., L. Zwarts, P. Callaerts, K. Norga, T. F. C. Mackay and R. R. H. Anholt (2008). "Neurogenetic networks for startle-induced locomotion in Drosophila melanogaster." Proceedings of the National Academy of Sciences of the United States of America **105**: 12393-12398.

Yamamoto, M. T. (2010). "Drosophila Genetic Resource and Stock Center; The National BioResource Project." Exp Anim **59**(2): 125-138.

Zanini, D., J. M. Jallon, L. Rabinow and M. L. Samson (2012). "Deletion of the Drosophila neuronal gene found in neurons disrupts brain anatomy and male courtship." Genes Brain Behav **9999**(999A).

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-829.

Zhang, Y., D. Sturgill, M. Parisi, S. Kumar and B. Oliver (2007). "Constraint and turnover in sex-biased gene expression in the genus Drosophila." Nature **450**(7167): 233-237.